NAVAL FACILITIES ENGINEERING SERVICE CENTER
Port Hueneme, California 93043-4370

TECHNICAL REPORT
TR-2083-OCN

ANALYSIS OF OCEAN WAVE FIELDS USING THE HARMONIC PHASE TRACKING PARAMETER ESTIMATION TECHNIQUE

by

P.A. Palo, Ph.D

September 1997

19971021 148

Approved for public release; distribution is unlimited.

Printed on Recycled Paper

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-018

Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
September 1997

3. REPORT TYPE AND DATES COVERED
Final 01 Oct 92-15 Jun 97

4. TITLE AND SUBTITLE

Analysis of Ocean Wave Fields Using the Harmonic Phase Tracking Parameter Estimation Technique

5. FUNDING NUMBERS

6. AUTHOR(S)

P.A. Palo, Ph.D

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESSE(S)

Naval Facilities Engineering Service Center
1100 23rd Avenue
Port Hueneme, CA 93043-4370

8. PERFORMING ORGANIZATION REPORT NUMBER

TR-2083-OCN

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESSES

Office of Naval Research
800 Quincy Street
Arlington, VA 22217-5660

Naval Facilities Engineering Service Center
1100 23rd Avenue
Port Hueneme, CA 93043-4370

10. SPONSORING/MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

Also published as Ph.D. dissertation, University of California, Santa Barbara, 1997

12a. DISTRIBUTION/AVAILABILITY STATEMENT

Approved for public release; distribution is unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

A uniquely new time series analysis technique called "Harmonic Phase Tracking" (HPT) was developed and used to examine the spatial and temporal evolutionary characteristics of a hurricane wave field. The fundamental motivation was to investigate whether ocean waves were random, or if they instead self-organize into a finite number of locally stationary discrete sinusoids.

Instead of the uniformly-spaced set of component frequencies inherent with the commonly-used Fourier Series (FFT) representation of a time series signal, HPT estimates the true number of harmonics along with the true frequencies, amplitudes and phases. HPT can be applied to wideband signals, and the parameters can be slowly-varying. Deterministic versus stochastic components are also readily identified.

This new HPT-based representation has great promise for the better understanding of ocean waves. Two sets of ocean waves were analyzed: a control group that corresponds to stationary conditions, and a second set that corresponds to Hurricane Bob. The analysis clearly showed that there is in fact a coherent and discrete structure to the energy content in the wave field; that is, waves do "self organize" into recognizable wave packets, with parameters that evolve very slowly over space and time.

14. SUBJECT TERMS

Signal processing, parameter estimation, harmonic retrieval, ocean waves, ocean engineering.

15. NUMBER OF PAGES
430

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	UL

EXECUTIVE SUMMARY

This study is best summarized as the development of a new signal processing technique followed by its engineering application to ocean wind waves. The fundamental motivation was, are ocean waves random, or do they self-organize into a finite number of discrete sinusoids? And if they do self-organize, could their behavior be better understood with a more realistic mathematical decomposition?

A uniquely new harmonic retrieval technique called "Harmonic Phase Tracking" (HPT) was developed. Unlike the usual Fourier Series representation which uses a uniformly-spaced set of component frequencies, HPT estimates the true number of harmonics (signal rank) along with their true frequencies, amplitudes and phases. A very simple fact serves as the basis for HPT: for a harmonic signal, it is possible to recover the true phase using only an estimated frequency. HPT exploits this by finding a series of true phases for shifted windows of the time series; the slope of the unwrapped phase versus time equals the true frequency. For the general case with multiple sinusoids (wideband signals), total least squares is used and iteration is required to converge to the best-estimated parameter set.

HPT is applicable to any time series and parameters can be slowly-varying. Deterministic versus stochastic components are identified from inspection of parameter evolution versus time. HPT is an adaptive, "high resolution" technique, meaning it is not a linear operator and that the frequency resolution exceeds the Rayleigh limit. Extensive validations are included for analytical and laboratory signals where the signal parameters are known.

This new HPT-based representation has great promise for the better understanding of ocean waves and any other signals with physically-

present sinusoidal components. Two sets of full-scale ocean waves are presented here. The first control group corresponds to stationary conditions, while the second set corresponds to Hurricane Bob on August 18 and 19, 1991. Both data sets include waves at multiple gage positions in 8 meters of water off Duck, NC. The results demonstrate that there is a coherent and discrete structure to the waves that evolves very slowly over space and time, as inferred from inspection of the evolution of mean frequency and amplitude versus time, and most importantly, from inspection of the continuous phase versus time for various components.

TABLE OF CONTENTS

	Page
1. INTRODUCTION	1
2. STOCHASTIC OCEAN WAVE ISSUES	9
3. REVIEW OF ESTIMATION TECHNIQUES	23
3.1 Background	23
3.2 Traditional Spectral Analysis	27
3.3 Subspace Estimation Techniques	45
3.4 Wavelet and Other Local Techniques	60
3.5 Chapter Summary	63
4. DESCRIPTION OF HARMONIC PHASE TRACKING METHOD	65
4.1 Chapter Overview	65
4.2 Algebraic Development of HPT as Applied to a Single Sinusoid	67
4.3 Geometric Interpretation of HPT as Applied to a Single Sinusoid	73
4.4 Review: HPT Phase Tracking and Frequency Correction as Applied to a Single Sinusoid Signal	78
4.5 Effect of Additive Noise in a Single Sinusoidal Signal	80
4.6 HPT Algebraic Development as Applied to a Multiharmonic Signal	83
4.7 Identification of the HPT Initial Frequency Vector for Multiharmonic Signals	98

5. VALIDATION OF HARMONIC PHASE TRACKING USING ANALYTICAL SIGNALS 113

5.1 Chapter Overview 113

5.2 Applicability of Harmonic Phase Tracking to Analytical Signals 117

5.2.1 Description of Multicomponent Analytical Signal 117

5.2.2 Representative Analysis of Multicomponent Signal 119

5.2.3 Representative Analysis of Single Sinusoid with Time Dependent Amplitude 134

5.2.4 Representative Analysis of Sinusoids with Time Dependent Frequency 141

5.2.5 Representative Analysis of White Noise Signal 159

5.2.6 Representative Analysis of Multiharmonic Signal with Additive White Noise 168

5.3 Numerical Aspects of Harmonic Phase Tracking 173

5.3.1 Data Segment Length and Bin Resolution 173

5.3.2 Investigation of Dependence of HPT Estimates on Initial Frequency Vector 186

5.4 Summary of Harmonic Phase Tracking Validation 189

6. ILLUSTRATION OF HARMONIC PHASE TRACKING USING PHYSICAL SIGNALS WITH KNOWN CHARACTERISTICS 193

6.1 Chapter Overview 193

6.2 Tidal Record Analysis 194

6.3 Rank 4 Laboratory Wave Signal 197

6.4 Pierson Moskowitz Laboratory Wave Signal 204

6.5 Frigate Heave and Wave Signals 210

6.6 Chapter Summary 214

7. HARMONIC PHASE TRACKING ANALYSIS OF OCEAN WAVE FIELDS 217

7.1 Chapter Introduction 217

7.2 Description of FRF Wave Data 218

7.3 Quasi-Stationary Wave Field Analysis 222

7.4 Hurricane Bob Wave Field Analysis 239

7.4.1 Overview of Storm 239

7.4.2 HPT Parametric Studies 242

7.4.3 HPT Investigation of Hurricane Bob Wavefield 275

7.5 Summary of Wave Field Observations 303

8. HARMONIC PHASE TRACKING: A PERSPECTIVE 305

8.1 Chapter Introduction 305

8.2 Signal Processing Issues with Harmonic Phase Tracking 306

8.2.1 Interpretation of HPT Estimates. 306

8.2.2 HPT Random Error Issues. 313

8.2.3 Linear Algebra Issues with HPT 316

8.3 Engineering Issues with Harmonic Phase Tracking 320

8.3.1 Numerical Implementation of HPT 320

8.3.2 Comparison Between HPT and FFT Representations 322

8.3.3 Engineering Applications of HPT 327

REFERENCES 335

APPENDIX A. USEFUL ALGEBRA FOR MULTIHARMONIC SIGNALS 339

A.1 General Algebraic Expressions for Two Summed Sinusoids 340

A.2 Instantaneous Frequency 349

A.3 Further Algebraic Studies

A.3.1 Amplitude Normalization 357

A.3.2 Quadratic Solution 358

A.3.3 More than Two Sinusoids 360

APPENDIX B. NUMERICAL STUDY OF ESTIMATED PHASE FOR MONOCHROMATIC SIGNAL 363

APPENDIX C. HARMONIC PHASE TRACKING ALGORITHMS 379

C.1 Overview 379

C.2 Basic Algorithm Assumptions 381

C.3 Construction of the **R** Transform Matrix 383

C.4 Estimation of the best two components from a row of **R** 387

C.5 Estimation of the best one component from a row of **R** 393

C.6 Adjustment of the R matrix to find the initial frequency vector estimate $\hat{f}^{(0)}$ 394

C.7 Estimation of the best frequency vector $\hat{f}$ via the Harmonic Phase Tracking technique 397

LIST OF FIGURES

Fig.	Title	Page
2.1a	Representative Ocean Waves (Hurricane Bob)	11
2.1b	Spectrum of Representative Ocean Waves (Hurricane Bob)	12
3.1	Spreading of Fourier Amplitudes with non-integer Harmonic Signal	38
3.2	Example Decomposition of Inverse Fourier Transform of Non-Integer-Period Sinusoid Used in Figure 3.1	40
3.3	Example of Singular Value Vectors for a Deterministic Signal Comprised of Two Sinusoids	52
4.1a	Minimum HPT Error Signal when Estimated Phase Equals the True Phase	75
4.1b	Non-minimum HPT Error Signal when Estimated Phase Does Not Equal the True Phase	76
4.2a	Least Squares (LS) Geometry	92
4.2b	Total Least Squares (TLS) Geometry	92
5.1	Time Domain Comparison of True and Estimated Multicomponent Analytical Signal	121
5.2	Bin (Frequency) Domain Comparison of True and Estimated Multicomponent Analytical Signal	123
5.3	Distribution of Normalized Residual Error versus Bin Number and Iteration Number for Multicomponent Analytical Signal	127
5.4	Error in Fitting Each Shifted Segment versus Iteration Number for Multicomponent Analytical Signal	129
5.5	Stationarity of Amplitudes versus Starting Time for Multicomponent Analytical Signal	130
5.6	Sample Convergence of HPT Parameters for Multiharmonic Analytical Signal	132
5.7	Example of Signal Extrapolation Using HPT Estimates	133
5.8	HPT Estimated Components versus Bin Number for Chirped Sinusoid with 10% Frequency Nonstationarity	145

5.9 HPT- and FFT- Estimated Component Evolution versus Bin Number for Sinusoid with Linear Frequency Nonstationarity 148

5.10a HPT-estimated Bin Numbers and Amplitudes versus Time for a Signal with a Linearly-Varying Frequency 150

5.10b HPT-estimated Phases versus Time for a Signal with a Linearly-Varying Frequency 152

5.11 HPT- and FFT- Estimated Component Evolution versus Bin Number for Sinusoids with a Constant and an Oscillatory Frequency 154

5.12a HPT-estimated Bin Numbers and Amplitudes versus Time for a 2-Component Signal with a Constant and an Oscillating Frequency 156

5.12b HPT-estimated Phases versus Time for a 2-Component Signal with a Constant and an Oscillating Frequency 157

5.13 HPT Estimated Components versus Bin Number for Independent Segments of a White Noise Signal 162

5.14 Evolution of Bin Number Estimates for White Noise Signal 164

5.15a HPT-estimated Bin Number and Amplitude Evolution for One Representative Component in a White Noise Signal 166

5.15b HPT-estimated Phase Evolution for One Representative Component in a White Noise Signal 167

5.16 HPT Estimated Components versus Bin Number for Independent Segments of Multiharmonic Signal with White Noise 169

5.17 HPT- and FFT- Estimated Component Evolution versus Bin Number for Two Sinusoids with Closely-Spaced Frequencies 180

5.18 HPT- and FFT- Estimated Component Evolution versus Bin Number for Three Sinusoids with Closely-Spaced Frequencies 182

5.19 HPT- and FFT-Estimated Component Evolution versus Bin Number for Three Sinusoids Showing Sawtooth HPT Behavior 184

6.1 Sample tidal record, summer 1996, Barlows Landing Beach Massachusetts 196

6.2a Rank 4 Wave Signal with HPT and FFT Estimates 199

6.2b HPT- and FFT-estimated components for Rank 4 Wave Signal 200

6.3 Example HPT-estimated Parameters for the Rank 4 Wave Signal 203

6.4	Programmed and Measured Amplitude Functions for U. S. N. A. Pierson Moskowitz Wave Signal	205
6.5	Measured Pierson Moskowitz Wave Signal	206
6.6	Sinusoidal Amplitudes for the Pierson Moskowitz Wave Signal	208
6.7	Comparison of HPT Component Evolution for the Pierson Moskowitz Wave Signal	209
6.8	Comparison of Amplitudes versus Bin Number for Wave and FFG Heave Signals Over Full Bandwidth	212
6.9	Comparison of HPT Component Evolution for the Wave and FFG Heave Signals	213
7.1	Gage Identifiers and Coordinates for FRF 8 meter Array	221
7.2	Sample Wave Data for Quasi-Stationary Analysis	223
7.3	Spectra for Quasi-Stationary Wave Data	224
7.4a	Raw FFT-based Amplitude Spectra for 0415-0425, September 13 1990	226
7.4b	Representative HPT Components for 0415-0425, September 13 1990	228
7.5	Consistency of Direct and Correlated HPT Estimates for Gage 211	229
7.6	HPT Evolution for Gage 131, 0415-0615, September 13, 1990	231
7.7a	Frequency and Amplitude Evolution for a Representative Stationary Packet	233
7.7b	Phase Evolution for a Representative Stationary Packet	234
7.8a	Frequency and Amplitude Evolution for a Representative Nonstationary Packet	236
7.8b	Phase Evolution for a Representative Nonstationary Packet	237
7.9	Significant Wave Height for Hurricane Bob	240
7.10	Representative Spectra over Half-Hour Intervals for Hurricane Bob	241
7.11a	HPT Component Evolution for 3 Different Analysis Lengths	244
7.11b	Detail of HPT Component Evolution for 2 Different Analysis Lengths	245

7.11c	Detail of HPT Component Evolution for 2 Different Analysis Lengths	246
7.11d	HPT Component Evolution for Doubled Analysis Length	247
7.12	Wave Packet Evolution prior to Hurricane Bob	249
7.13a	Representative Wave Packet Evolution prior to Hurricane Bob	250
7.13b	Phases of Representative Wave Packet prior to Hurricane Bob	251
7.14	Phases of Representative Wave Packets for Doubled Segment Analyses prior to Hurricane Bob	253
7.15	Comparison of HPT and FFT Component Evolutions during the initial stages of Hurricane Bob	254
7.16	Representative In-line and Orthogonal Correlation Functions prior to Hurricane Bob	256
7.17	HPT Estimates for Gages at 130m spacing	257
7.18a	HPT Component Evolution over Main Bandwidth of Energy for Gages 251 and 211 prior to Hurricane Bob	258
7.18b	HPT Component Evolution over Secondary Bandwidth of Energy for Gages 251 and 211 prior to Hurricane Bob	259
7.19	Comparison of Equivalent Wave Packets for Gage 251 and Gage 211	261
7.20	HPT Component Evolution over Main Bandwidth of Energy for Gages 131 and 191 prior to Hurricane Bob	262
7.21	Representative Wave Packets for Gage 131 and Gage 191 prior to Hurricane Bob	263
7.22	Illustration of Raw Gage Frequency Vectors and Best, Mean Frequency Vector	266
7.23	Example Adjusted and Fitted Phases for Frequency = 0.148 Hz	269
7.24	Estimated and Analytical Wavelengths versus HPT Segment Length, with FFT Reference	271
7.25	Detail from Figure 7.25a of Estimated Incident Directions	273
7.26	HPT-Estimated Incident Directions (Deg) versus Frequency (Hz) During Hurricane Bob	276

7.27	Wave Packet Evolution during the peak of Hurricane Bob for Gages 111, 251, and 191	277
7.28	Representative Packets During Peak of Hurricane Bob	279
7.29	Wave Packet Evolution at Higher Frequencies for Gages 251 and 111 near Peak of Hurricane Bob	281
7.30a	Low Frequency Wave Packets for Gages 251 and 211 after Peak of Hurricane Bob	283
7.30b	Peak Frequency Wave Packets for Gages 251 and 211 after Peak of Hurricane Bob	284
7.30c	Above Peak Frequency Wave Packets for Gages 251 and 211 after Peak of Hurricane Bob	285
7.31	Illustration of Disappearing Wave Packet after Peak of Hurricane Bob, Gages 251 and 211	286
7.32a	Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak in a Stationary Wave field	289
7.32b	Illustration of Spatial Homogeneity for Frequencies Across the Spectral Peak in a Stationary Wave field	290
7.32c	Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak in a Stationary Wave field	291
7.33a	Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak During the Initial Stages of Hurricane Bob	293
7.33b	Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak During the Maximum Stages of Hurricane Bob	294
7.33c	Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak After the Maximum Stages of Hurricane Bob	295
7.34a	Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak During the Initial Stages of Hurricane Bob	296
7.34b	Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak During the Maximum Stages of Hurricane Bob	297
7.34c	Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak After the Maximum Stages of Hurricane Bob	298

7.35	Example Clusters Showing Spatial Homogeneity	299
7.36a	Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak During the Initial Stages of Hurricane Bob	300
7.36b	Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak During the Maximum Stages of Hurricane Bob	301
7.36c	Illustration of Spatial Homogeneity for Frequencies above the Second Spectral Peak After the Maximum Stages of Hurricane Bob	302
8.1a	Representative HPT Signal Extrapolation Relative to Center Segment Using Waves from Hurricane Bob	328
8.1b	Representative HPT Signal Extrapolation Relative to Stationary Frequency Vector Using Waves from Hurricane Bob	329
8.2	Wave Packet Evolution, September 13 1991, Gage 211	333
A.1	Illustration of Phase Discontinuity in Beating Sinusoids	344
A.2	Example of Phase Discontinuity (at 150 sec) in an Ocean Wave Signal (Prior to Hurricane Bob, Gage 131, at 1945 on Aug 18 1991)	345
A.3	Illustration of Instantaneous Frequency for Two Sinusoids	354
A.4	Illustration of Instantaneous Frequency for Three Sinusoids	355

LIST OF TABLES

Table	Title	Page
5.1	Constant Parameters for Multiharmonic Analytical Signal	118
5.2	HPT-Estimated and True Parameters for Multicomponent Analytical Signal	125
5.3	HPT Analyses of 3-Component Deterministic Signal versus Segment Length and Envelope Phase	176
5.4	Estimated Parameters for Multicomponent Analytical Signal Using Modified Initial Frequency Vector	187
6.1	Comparison of Astronomical and HPT-Estimated Tidal Periods for Figure 6.1	195
7.1a	Coordinates for North-South Array FRF Gages	219
7.1b	Coordinates for East-West Array FRF Gages	220
7.2	Environmental Parameters for Quasi-Stationary Waves	222
7.3	Example Directionality Estimate For Frequency = 0.148 Hz	268
7.4	Gage Nomenclature for Homogeneity Figures	288
8.1	Exact HPT Estimated Bin Numbers and Amplitudes for a Square Wave Signal	324
B.1	Bias Errors in the Phase Estimates	369
B.2	Bias Errors in the Estimated In-Phase Coefficients	372
B.3	Bias Errors in the Estimated Out-of-Phase Coefficients	375
C.1	Appendix C Outline	383

CHAPTER 1

INTRODUCTION

Ocean engineering can be described as the design of physical objects that operate in large bodies of water. It is a diverse field of study that encompasses the disciplines of civil, mechanical and electrical engineering, as well as chemistry and physics. Sample applications include ships, buoys, breakwaters, piers, submersibles, moorings, oil exploration platforms, propulsion, coastal erosion, and environmental instrumentation. For the majority of these applications, the key to their optimal operation and/or survival is understanding their interactions with various types of surface waves. It is not surprising, therefore, that the study of waves and how they dynamically excite these objects constitutes the greatest single emphasis in ocean engineering research and design.

Decades of intense research and measurements on ocean waves have provided ocean engineers with a reasonable capability to confidently

place objects in the ocean such that they will perform their intended function at an economical cost. However, it cannot be inferred from this statement that ocean engineers have a "reasonable" capability to locally describe a typical wind-driven wavefield either spatially or temporally. At one extreme, there are global descriptors such as probability density and spectral functions that have proven useful. At the other extreme, it is sometimes possible for design purposes to approximate the irregular surface as a single monochromatic wave using small or finite amplitude wave theory with a stochastically averaged amplitude and period. But neither of these approaches are acceptable for a great many problems, and neither provide any level of insight into the wavefield itself. As one example, the widely used linear wave theory only describes the wave *below* the mean free surface, and this creates a fundamental conceptual problem regarding the dynamic velocity field for arbitrary bichromatic waves. If the frequencies are approximately equal, then the velocity field is consistent (both fields referenced to the mean free surface) and straightforward to interpret. But if the frequencies are widely separated such that there is a high frequency wave is superimposed on a much lower frequency wave, the velocity field physically corresponds to the higher frequency monochromatic wave superimposed on what looks like a quasi-static mean free surface with a time-varying current (e.g., the lower frequency monochromatic wave). In other words, the velocity field for one of the two bichromatic waves is now referenced not to the mean free surface as in the first case but to the instantaneous surface of the

second wave - even above the mean free surface. The transition between these two cases is not clear. As a second example, consider the superposition of two collinear waves in an intermediate water depth. It is possible to have two waves with different wavelengths but the same frequency if the depth is such that it produces Stokes harmonics for the lower frequency shoaled wave but not for an independent free wave at one of the super harmonic frequencies. In this case the Stokes harmonic wavelength is an (even) integer harmonic of the fundamental low frequency wavelength which does not obey the dispersion relationship. As before, the mathematical treatment of this case is unclear.

This dissertation addressed the need for an analysis tool to allow for a physically meaningful local descriptor of a wave field. In essence, such an analysis tool would provide suitable information to "bridge the gap" between the examples described above.

An entirely new technique called *Harmonic Phase Tracking (HPT)* is presented. This parameter estimation technique is not restricted to any signal or noise characteristics such as rank or normality. HPT decomposes a finite length deterministic or stochastic signal into a set of discrete sinusoids of minimal rank with constant but arbitrary amplitudes, phases and frequencies (subject to certain time-frequency ambiguities), with noise that is not necessarily white nor uncorrelated. It is demonstrated that sinusoids physically present in the signal can be confidently

identified using ensemble averaging and phase continuity. This is a significant new capability offered by HPT, and it is used to provide a wealth of new information for a wide range of physical signals, particularly mildly nonstationary time series where time-frequency distributions would only provide approximate qualitative and quantitative insights.

At one extreme, HPT is best applied when there is reason to believe that the signal is multiharmonic with a discrete frequency vector - for example, with some geophysical processes, structural responses, radar, sonar, etc. Conversely, it is also shown that the inherent ability of HPT to optimize a frequency vector means that it can also efficiently and reliably model and identify even white noise. Subsequent post-processing analyses are then possible to estimate other vector parameters of engineering interest such as stationarity, homogeneity, and directionality. The major emphasis in this ocean wave study involves development and validation of the signal processing technique itself. But given the powerful new capabilities offered by HPT, that emphasis is understandable, and is not that uncommon when compared to the development of many other engineering analysis tools (for example, the numerical integration techniques vital to diffraction theory).

The second chapter presents some of the issues related to the description of typical ocean wave fields. There are two conclusions: (1) that there is strong evidence that a slowly-varying spatial and temporal structure to

typical wave fields does exist, but (2) that there are no existing techniques available for identifying the structure. Chapter 3 reviews the strengths and limitations of available spectral analysis ("low" resolution) and parameter estimation ("high" resolution) tools. Spectral analysis is discarded as a non optimum choice for quantifying a wave field for a variety of reasons, primarily: (1) the fact that the orthogonal Fourier harmonics are not physically meaningful, (2) biasing (leakage) between Fourier bins, and (3) the need for ensemble averaging to get statistically meaningful results. Similar arguments hold for wavelet analyses. The high resolution techniques are likewise discarded because they assume white noise and can only be applied to low rank signals.

Chapter 4 presents the fundamentals of the new Harmonic Phase Tracking (HPT) parameter estimation methodology that was developed in this study. The key fact exploited by HPT is that a least squares fit between a monochromatic signal with *unknown* parameters, and a trial monochromatic signal with only an *approximately-correct frequency*, allows for recovery of the *true* phase. Then, if several [true] phases are identified for a series of segments shifted in time, the slope of this phase vector versus time then equals the *true* frequency. The extension to multiharmonic signals follows naturally as an iterative analog. The HPT model is a very "natural" basis consisting of the minimum number of sinusoids, with rank and frequencies dictated by the signal properties, needed to match the time domain signal (with some exceptions). It is a

high resolution, adaptive parameter estimation technique that can be applied regardless of the (unknown) signal rank. Ensemble averaging of successive HPT parameter estimates (evolutionary plots) allows it to be used quite effectively as a time-frequency analysis tool for nonstationary processes.

Chapter 5 is a very long and thorough Chapter that presents benchmark validations using a variety of *analytical* (deterministic) and stochastic signals including: multiharmonic signals with and without noise; nonstationary sinusoids, and chirped sinusoids. Some of the Chapter conclusions include demonstrations that: (1) HPT finds the correct rank with exact amplitudes, frequencies and phases (is unbiased for deterministic signals); (2) noise biases estimated signal amplitudes but signal frequencies are relatively unbiased; (3) noise can be easily identified and eliminated with ensemble averaging ; (4) the technique is insensitive to linear amplitude variations; (5) it does model sinusoids with varying frequency; (6) time-frequency resolution approaching half the Fourier value is (conservatively) achieved; and (7) HPT is numerically robust. In short, this new "Harmonic Phase Tracking" technique is a valuable new contribution to the signal processing community in general, and it is shown to accommodate the non constant amplitudes and frequencies, high rank, and unknown noise characteristics expected for typical ocean waves.

Chapter 6 investigates *physical water wave signals with approximately-known parameters*. This includes ocean tide and laboratory scale forced waves from a wave generator. In each case HPT either confirmed parameter values (like tidal periods) or suggested more accurate parameters (e.g., HPT estimates imply that one particular wavemaker has internal resonances such that it produces frequencies that deviate from the equally-spaced Fourier set).

Chapter 7 is the main focus of this study, and it addresses *ocean waves* using data measured at an extended bottom-mounted array of gages. Wave fields are analyzed representing both stationary and nonstationary (Hurricane Bob) conditions. This multi-gage array allows for spatial (homogeneity) as well as temporal (stationarity) checks on the variability of the estimated stochastic parameters. HPT allows for new insights into ocean wave fields that are not based on inspection of averaged parameters from existing orthogonal (i.e., nonphysical) techniques. Conclusions and observations from this Chapter include: (1) ocean waves do "self organize" into discrete packets that propagate coherently for time scales approaching one hour (i.e., it is not necessary to assume the spectrum is continuous); (2) in general, the frequencies of the discrete packets decrease slowly versus time; (3) multiple packets can merge and, single packets can split; and (4) packets at widely-separated gages show a high degree of spatial coherence as measured by their respective frequencies, amplitudes, and phases.

Chapter 8 presents a perspective on HPT in terms of: error measures and interpretation issues, HPT linear algebra and algorithm improvements, and further engineering applications.

The target audience for this dissertation is not signal processing analysts, but rather the much larger field of practicing scientists and engineers who use (and are often overly trustful of) spectral analysis and other parameter estimation techniques as tools in their respective fields of study. For this reason each Chapter features qualitative descriptions and straightforward examples intended to clearly demonstrate the significance of HPT results versus other widely-used results.

Standard nomenclature is used throughout. Non-bold lower case letters are either scalars (e.g., f or f_1 is frequency) or continuous functions (e.g., t or x(t)). Lower case bold letters are (typically time domain) column vectors (e.g., **x** represents a discrete time series). Upper case bold letters are either matrices or frequency domain transforms of the same lower case time domain vector (e.g., **X** is the Fourier Transform of **x**). Variables with subscripts denote individual values from a vector or matrix (e.g., x_k, X_k, or A_{ij}). Greek letters are scalars, often phase. Variables are defined where first used.

CHAPTER 2

STOCHASTIC OCEAN WAVE ISSUES

Who actually needs an accurate vector description of a general ocean wave field? First, consider the designer of an ocean-based facility such as a compliant moored floating production system in relatively deep water. This particular type of facility has a variety of resonant phenomena such as first-order ship and cable responses and second-order drift motions. Design of such a facility requires accurate knowledge of the instantaneous waves and the corresponding wave envelope in both time and space. While there are a host of other design issues such as life cycle costs, maintenance and inspection, etc., the single most important analysis objective is to quantify the system responses to operational and survival wave environments. As a second example, consider a physical oceanographer studying the mechanisms of energy transfer in a wind-wave field. How does the primary frequency evolve in time and space? Under what conditions are harmonics (sidebands) created? How do superimposed waves evolve? Spectral functions resulting from long

ensemble averages do not provide the level of detail needed to accurately identify the relevant processes.

Assume that a limited set of wave measurements exists that will be used to define these or other wave fields. What questions are natural for an engineer or scientist to ask to quantify them? What parameters/functions would most *ideally* describe a wave field? Are they practically obtainable using available numerical techniques? How do the limitations and uncertainties in these obtainable descriptors affect other derived estimates such as directionality? Are statistically-averaged descriptors always adequate?

It is instructive here to physically and qualitatively describe the characteristics of "typical" ocean waves. A wave field often consists of 2 or more sub fields propagating from different directions over different frequency bands. The dominant sub field is typically comprised of waves between 1 and 12 seconds (0.08Hz to 1Hz). These waves are generated locally (e.g., within a hundred mile radius) by the wind, with energy propagating in a mean direction but exhibiting significant spreading. This spectral band is typically narrowband, with pronounced grouping evident in the time history realizations. Superimposed with this wind-generated wave field is (at least) one independent lower-frequency sub field which usually has much less energy than the local wave field. These swell waves typically have periods greater than 18 seconds (0.055Hz), and

are the remnants of a wind wave sub field after traveling from a very distant (thousands of miles) generating region. Figure 2.1a shows a representative time history of one typical wave field, with the corresponding spectrum shown in Figure 2.1b. Note first that while the primary bandwidth of the spectrum is narrowbanded, the waves do contain energy over a wide band of frequencies. The wave signal was low-pass filtered to isolate the low frequency swell subfield (centered around 0.02 Hz in Figure 2.1b); this component is superimposed as a dotted line in Figure 2.1a. Energy above the primary spectral bandwidth corresponds to a distinct third wave regime representing wind waves in the early stages of growth.

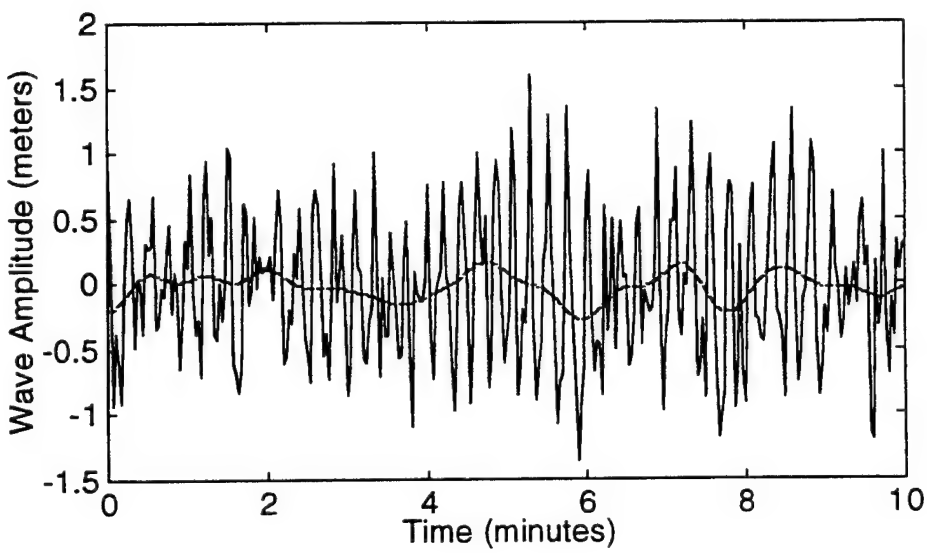

Figure 2.1a Representative Ocean Waves (Hurricane Bob)

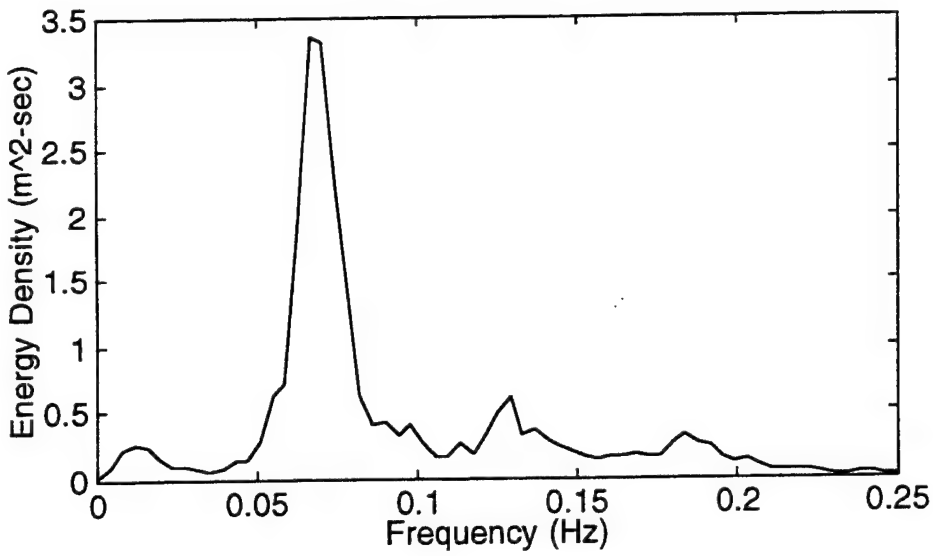

Figure 2.1b Spectrum of Representative Ocean Waves (Hurricane Bob)

These figures are presented merely as representative of the wide variety of wave fields that are possible. Note that these figures and this study do not address other categories of waves such as seiche and capillary that are present but not significant compared to the local and swell waves. Second, this study will purposely avoid extreme shallow water waves where nonlinear superharmonics are bound to the primary frequencies. Medina and Hudspeth (1990) reference 4 previous studies which concluded that the linear hypothesis, that is, that the waves propagate independently, is valid in water depths greater than 10 meters, so we can take that as an approximate defining measure of deep and shallow waves.

What additional qualitative descriptors can be added regarding ocean wave fields? It seems safe to start by confidently stating that the waves are

definitely nonperiodic at any time scale. From a mathematical perspective, this requires that the spectrum must be continuous to insure that there are no harmonics that would be periodic - IF any such components were physically present in the signal. As discussed later, this condition may be relaxed depending on the time scale.

Is a typical wavefield ever stationary? Strictly speaking, no. However, to answer this correctly we need to define two things: a quantitative measure and a time scale. It is prevalent but not always justifiable in many engineering studies to conveniently select either or both such that the wavefield is assumed to be "stationary". Various measures of stationarity include: strict (or isotropic); weak (or wide-sense); strong (or, completely, or strict-sense); or stationary to order K (Bendat and Piersol, 1986; Trevino, 1982; Jenkins and Watts, 1968; Papoulis, 1965). Usually, a signal is considered stationary if the second order statistics (K=2, for mean and standard deviation) are invariant with time (i.e., weakly stationary).

The next issue is the time scale. Defining the time scale of interest actually controls whether the question of stationarity is even tractable. At one extreme, waves observed over a span of several minutes can appear fairly uniform in amplitude and period, even after decomposition into finite frequency bands. At the other extreme, if the goal is to describe the evolution of the waves before, during, and after a long multi-day storm, then the frequency resolution must be so small as to make the spectrum

essentially continuous. Accordingly, ocean engineers and oceanographers typically define a time span of interest of no more than several hours as sufficient "for engineering purposes". Elgar and Seymour (1985) concluded that taking 17 minutes of data 4 times a day provided adequate estimates for significant wave heights (proportional to the standard deviation and therefore implying weak stationarity) compared to longer estimates based on records of typically 5 hours; in other words, sampling every 6 hours was sufficient. On the other hand, Toba, et al (1988) concluded that for growing seas the time scale was more like "10 to 15 minutes to produce a shift in the peak frequency." This time scale is consistent with assumptions from Goda (1985) and Athanassoulis, et al, 1992. Detecting whether the slow shifts of statistical (often scalar) descriptors relevant to nonstationarity are significant is not detectable by simply tracking the standard deviation (scalar area under the spectrum) - demonstrating that tests for "weak" stationarity are at best approximations for the true conditions. Ideally, if the signal could be modeled locally as a finite number of sinusoids, then a more absolute measure of stationarity would be to examine the variation in *all* of the component parameters versus time.

The next natural question is complementary to the temporal stationarity issue and relates to the spatial homogeneity of the wave field. This can be more important than the question of stationarity, particularly in the design of structures with dimensions at least as large as the wavelengths;

one example is the proposed U. S. Navy Mobile Offshore Base (MOB) which has a length scale of approximately 1.5 km. What are the typical spatial dimensions associated with waves both parallel and perpendicular to the direction of wave advance? The latter is often labeled "short crestedness". For this, we can postulate that the answer would vary depending on the circumstances of the waves (fetch, wind stationarity, etc.). One quantitative engineering answer to this question was given by Hughes (1986) who processed wave measurements from an offshore array to present a plot of "maximum coherence" (with and orthogonal to the wave advance) versus gage separation normalized by the wavelength. He concluded that individual waves are indeed typically "short crested" as measured by the coherence over distances of one or at most a few wavelengths. So, this result does give us one quantitative measure with respect to one distance scale (approximately one wavelength).

It is also instructive to invoke philosophical arguments to help spatially and temporally describe a wave field. Consider, for the moment, a wavefield that is not actively growing from the wind. It is known that the energy continually spreads over a range of directions relative to the principle wave advance, and that this spread decreases with propagation distance, resulting in more long crested waves (Kinsman, 1984). This spreading is further reduced as the water depth decreases (Sorensen, 1993). Assuming that if a reasonable (not quantitatively defined) propagation distance existed, we would expect fairly uniform conditions

based on philosophical arguments such as conservation of energy flux for a unit width of the wave. This is, in fact, the justification for ray tracing with wave refraction studies. Furthermore, many investigators formally argue that the energy in each frequency band of the spectrum should be only a "slowly-varying" (relative to the peak period) function of space and time (Komen, et. al., 1994; Kinsman, 1984). Since this implies that the energy in each frequency band does propagate smoothly, at least over short distances, it therefore seems reasonable to propose a model for a finite segment of stochastic ocean waves consisting of a finite summation of harmonic waves (with an as-yet unknown discrete frequency set and amplitudes with an unknown spatial dependence). These narrowband segments that are coherent over slowly-varying time scales are often referred to as "wave packets". Note also that the use of Fourier Series is only a crude approximation to this finite, discrete summation because of the finite frequency resolution and the condition of periodicity implicit in the approach (discussed further in the next Chapter).

There is a third approach for addressing this issue of homogeneity, and that is physical observations. Kinsman (1984, p543) offers an excellent quote next to an aerial photograph of waves:

> *The sea surface is irregular but, viewed from above, it also seems to have considerable regularity. Notice particularly the fairly regular arcate structure strongly marked by the breaking crest to the right of center. Mr. R. G. Stevens, W.H.O.I., first drew this regularity to my attention. He tells me that he has found that*

when the regularity of the pattern in a growing sea is apparently lost it can often be seen again if the sea is viewed from a higher level. This suggests that neither purely deterministic nor purely stochastic mathematical models are likely to be adequate for an understanding of the physics of the sea. The regularity indicates that some selective process may well be at work even in the later stages of the growth of a sea."

In other words, while the entire field is undeniably stochastic, there is evidence of deterministic behavior at some intermediate spatial scale.

Radar and stereo-photogrammetry measurements loosely qualify as "physical observations" and fall into this category. Here, too, structures can be observed in apparently "irregular" seas, typically at a scale proportional to the wave groups (or packets). Werle (1996) illustrates the intensity of a low grazing angle radar backscatter signal versus range and time, and concluded that the dimensions of the most apparent features were proportional to wave "packets" or "groups" that were evidently quite organized. Stereo-photogrammetry, while very difficult and expensive to apply, also yields contour mappings of wave fields that allow for ready identification of spatial characteristics (Goda, 1985; Horikawa, 1988).

In a broad sense, these visual and radar observations are consistent in finding homogeneity and structure at a scale slightly larger than the apparent waves. That scale of structure is likewise consistent with the previously described philosophical expectation that amplitudes and phases

should be "slowly varying", that is coherent or organized, over large distances and time. However, neither observational technique is capable of decomposing the wave field into its constituent components (however, in the limit of a very narrowbanded process the waves approach regularity and in these cases of essentially one wave component these techniques can be useful).

While the discussion over the last several pages does not answer the question of spatial homogeneity, it does establish that it is suspected to exist over finite-sized regions of a wavefield and is a worthwhile question to address in research.

Another necessary parameter for describing a wavefield is the number of distinct wave directions simultaneously present - whether the sea is uni- or multi-directional. Note that an array of multiple wave gages is required to resolve this question, subject to two assumptions when estimating directionality:

- all components in a given frequency band are unidirectional; that is, incident energy bands are non overlapping, and
- the scale of short crestedness is a known function of frequency.

The first condition of non overlapping bands can be practically satisfied when no more than one swell and/or one locally wind-generated wave field are present. However, this is typically not true for the more common

case of multiple wind-generated wave fields. In these latter cases, the complex Fourier Series ordinates at each bin from one subrecord will physically correspond to the average amplitude and phase of all the component waves (0 to Nyquist frequency) and all wave directions correlated with that Fourier frequency. Thus, there is large uncertainty associated with each estimate. It is well known that corresponding cross-spectral phase functions (used to estimate directionality) are useful only if a large number of subrecords are averaged.

The second condition regarding short crestedness cannot be accurately evaluated. Since there is no existing technique for determining the length scale for how short crested a wave field really is, a conservative length is assumed - for example, one wavelength at a frequency of interest (which agrees with the conclusions from Hughes, 1986). Then, cross spectra/correlations between the finite-spaced gages and the wave dispersion equation are used to estimate directionality. [The array dimensions therefore predetermine which frequencies can be considered.] In cases where the actual short crestedness is less than the assumed value, then cross spectra may still asymptotically estimate the correct incident direction, but the variance will be much larger and, again, a large number of ensembles will be required to account for that uncertainty. Thus, knowledge of the general spatial structure of the wavefield, such as the scale of short crestedness, is presently unavailable but inherently important in estimating directionality from a given array

because at the very least it establishes the minimum amount of statistical averaging required.

One last question can be asked regarding the probability distribution for the instantaneous wave amplitudes. This seems like a relatively straightforward question to answer using a histogram of available data. However, it indirectly is involved with several other issues and is therefore important. The first issue that comes in is ergodicity of the wavefield. A nonlinear wave in very shallow water has a definite non-Gaussian distribution, as measured over space at a fixed time or at a fixed position versus time. However, an ensemble average at many gages at a fixed time would appear Gaussian - thus making a very shallow wavefield technically nonergodic (Tucker, 1995). Second, it is known that if a linear (i.e., not shallow water) wavefield is both stationary and homogeneous, then because of the dispersive nature of the waves the sea will be comprised of a large number of independent sinusoids and therefore the distribution must be Gaussian (Komen, 1994). Both of these conclusions are seen to be violated if wave coupling - nonlinearity - is present (and, conversely and just as importantly, they are not violated for linear conditions). Such coupling is best modeled as two phase-locked sinusoids at integer-multiple frequencies (e.g., Stokes components) - illustrating yet another example of "deterministic behavior" in a stochastic wave field. While this study does not address such shallow water waves, this does further demonstrate the potential value of a tool that would allow for

identification of discrete vector sinusoids from a random wave record. The bispectrum is presently used for this type of study but it has two drawbacks: (1) it is FFT-based so it suffers from low frequency resolution, and (2) the uncertainly associated with bispectral estimates typically requires an order of magnitude more data than the usual spectrum, which is a serious problem for ocean waves because of stationary limitations.

What is the final verdict of all of these discussions? There are two. First, there indeed does seem to be an as-yet unidentified spatial and temporal structure to a typical irregular wind-driven sea that is worthy of study. A final quote from Kinsman (1984, p402) is again instructive:

> *Our common sense tells us that, even within a generating area, any closely suitable spectrum must be a slowly varying function of space and time. The process simply cannot be a* stationary Gaussian *process as we have been assuming. This is most unfortunate. You may rest assured if I had any idea of how to build a manageable, nonstationary, Non-Gaussian model of a stochastic process that would be a better approximation to the sea surface, I'd do it. Fortunately, depending on the weather, over areas of hundreds of square miles and for many hours at a time, wave records lasting from a few minutes to several hours often look so much like cuts from a stationary Gaussian process that you might just as well treat them as thought they were. The differences are unimportant. Such a process, whose statistics will remain invariant under time and space translations up to a certain size, we call* quasi-stationary. *If you insist on being very precise in your language, you will describe the sea surface as a* quasi-stationary, pseudo-Gaussian *process."*

Kinsman confirms the main conclusion of this Chapter that there should be some structure in a wavefield, and that the component spectral amplitudes should be *slowly varying* functions of time and space (i.e., quasi-stationary). He then suggests the usual approach to wave analysis - to assume local stationarity and homogeneity even though he knows better (although it is not stated exactly which statistics he feels will be invariant). But most importantly, Kinsman directly confirms the thrust of this study when he states the need for a better stochastic model. In doing so he indirectly confirms the second conclusion from this section, namely, that existing techniques have not been successful in satisfactorily answering these questions about ocean wave fields.

Those techniques are evaluated in the next Chapter.

CHAPTER 3

REVIEW OF ESTIMATION TECHNIQUES

The field loosely defined as spectral analysis has long been described as an "art as much as a science." And, while there has been a virtual explosion of new techniques during the last two decades that deliver revolutionary improvements in resolution, it cannot be denied that results are still very dependent on the insight and skill of the analyst. This can be just as frustrating to the analyst generating the results as it is to the engineer or scientist who needs to use the results. The objective of this Chapter is to review the assumptions, concepts, and consequences of available signal processing techniques as they may apply to the analysis of ocean waves.

3.1 Background

The tools and terminology of signal processing can be overwhelming to anyone who is not an active practitioner in the field. Terms like Nyquist frequency, Hamming and boxcar windows, ARMA models (with zeros and

poles), Maximum Entropy, ensemble averages (versus Expected Values), signal subspaces, Cramer-Rao bounds, degrees of freedom, leakage, and white versus colored noise quickly become such an alphabet soup that the significance of the final results are completely lost, or worse, misinterpreted. Obviously, this situation should be avoided at all costs. It is a very real danger for interdisciplinary studies like the present one, because the *engineering* objective of describing stochastic ocean waves can only be achieved by using and understanding the *mathematical* language of stochastic signal processing.

Naidu (1996, p347) states this extremely well:

> *"Perhaps the most difficult question in modern spectrum analysis is,* how does one determine the right model of a time series whose finite duration sample is available as the observed data? *As of now, we do not have any clear-cut answer to this question; we have to depend upon the physics of the problem to surmise on the nature of signal and noise sources."*

Usually the first decision to make in signal processing regards the *a priori* assumptions about the underlying *signal* - in this case, ocean waves. There are advantages and potential disadvantages to doing this. On one hand, incorporating correct insights gives the analyst added flexibility in choosing the "best" technique for interpreting the measurements. However, if those insights are unfounded, then the results can be very misleading. For example, the final spectrum defined by the ensemble

average of a set of finite Fourier Transforms (FFTs) is meaningful only if the signal is known to be stationary. If we have no idea of whether the signal is stationary, or alternatively what measure of stationarity it possesses (weak, strong, etc.), then it would be necessary to resort to general time-frequency techniques such as Wigner-Ville distribution, Evolutionary Spectrum, etc. (see Boashash, 1992). If we suspected that the signal indeed was reasonably stationary (say, wide sense) but did not know whether the spectrum was continuous or discrete, then we could choose a nonparametric approach using a spectrum based on the FFT-based approach described above. However, we should not expect too much information other than "the approximate distribution of variance versus frequency" from that analysis. Techniques that require only that the signal be stationary represent the "lowest" level of modeling with respect to constraining the answer to fit any particular set of *a priori* conditions.

Conversely, we may instead have physical reasons to expect that the signal is comprised of a finite number of sinusoidal basis vectors resulting in a discrete "line" spectrum (for example, the first few mode shapes for the response of a multi-story, lightly-damped building are sinusoids with integer multiples for the periods). In this case the signal is the "highest" level of modeling - namely, a finite summation of linearly-superimposed sinusoids (with noise). If this is truly the case then many of the new parametric "subspace" methods described in this Chapter are appropriate. But it bears repeating that caution must be used because it is easy to over

specify the signal model and subsequently make untrue conclusions regarding the signal content. In other words, results which are tailored to *a priori* assumptions may or may <u>not</u> tell us whether that signal structure (as defined by the assumptions of that technique) is physically present, only what its structure would be *if* it was present.

The quote from Naidu also specifically acknowledges the need to at least qualitatively model the noise, which is defined as all effects not included in the signal model (instrumentation effects, nonstationarities, nonlinearities, etc.). It is very easy to overlook this step. Since there is usually very little insight available for quantifying this "noise", the assumption is often made for convenience that the level of the error is unknown but equally likely at all frequencies - precisely (bandlimited) white noise. In fact, as explained later in this chapter, the new subspace methods generally require this assumption. *In the present study of stochastic water waves this may not be true* - for example, it is equally plausible that "unknown effects" at the leading edge of a spectrum (like energy downshifting) or at the peak frequency (where energy is greatest and the chance of wave breaking is highest) are relatively larger than at the other frequencies, negating the white noise assumption and eliminating the use of subspace methods. This illustrates how serious consideration of just one facet of the analysis - in this case the unknown "noise" effects - could, by itself, steer the signal processing in an entirely different direction than anticipated.

Regardless, Naidu's ultimate advice is to depend on the physics of the situation (and not the convenience of the mathematics) to guide the signal processing. That will be done in the next chapter for this study of stochastic ocean waves. The remainder of this chapter focuses on two sections, one describing traditional (fast being replaced by the term "low resolution") spectral analysis techniques, and a second section describing the new subspace decomposition (or "high resolution") techniques; see Kay and Marple (1981) for an excellent tutorial. The information in these two sections is essential for understanding the scope of this wave study and the corresponding results. A brief third section is included that primarily addresses the application of alternative approaches (e.g., wavelets) to ocean wave analysis.

As anyone who has ever studied signal processing can attest, there is a very extensive range of mathematics associated with it. Not all of the basic theoretical concepts are addressed in this review. The theory of Fourier integrals and series for signals with finite and nonfinite energy is a good example, where measure theory is required to properly handle various types of signals (Carslaw, 1950 and Billingsley, 1986).

3.2 Traditional Spectral Analysis

Since the subspace techniques presented in the subsequent section have only been fully developed over the last decade or so, the techniques in this

first category are probably the most familiar to engineers and scientists. And, understanding their performance sets the baseline by which the new subspace techniques are compared.

Approaches that are classified as "traditional" are described in a great many texts and include:

- Blackman-Tukey spectrum based on the transform of the autocorrelation function using the Wiener-Khinchine relationship;
- Fast Fourier Transform (FFT) based spectrum using ensemble averaging;
- AR (autoregressive), MA (moving average) and ARMA spectrum based on polynomial fitting, and
- Maximum Entropy spectrum.

Since these can all be shown to be equivalent, assuming only that the spectrum is the transform of the autocorrelation (Naidu, 1996; p71, 242), for the purposes of this section only concepts inherent to the FFT-based approach will be detailed. This choice also provides useful background information for development of the new technique in Chapter 4.

The equation defining the FFT-based spectrum appears straightforward (e.g., Bendat and Piersol, 1986; Kay, 1988). Start with discrete samples of a continuous time series x(t) as $x_m = x(m\Delta t)$, m=1,2, ... L. Identify the [trial]

length (M) of the Fast Fourier Transform (FFT), then define the spectrum as:

$$\hat{S}(f_i) \equiv \lim_{N\to\infty} \frac{1}{N} \sum_{k=1}^{N} \left[\left(\frac{2}{T}\right) \left|\hat{X}_k(f_i)\right|^2 \right] \qquad 3.1$$
$$= E\left[\tilde{S}(f_i)\right]$$

where: $\hat{S}$=spectral estimate at discrete frequency f_i (discussed momentarily), $N = \text{int}(L/M)$ = integer number of independent ensemble averages, $T=M\Delta t$ is the length of the FFT integration interval, $\hat{\mathbf{X}}$=complex Fourier transform of the finite length vector **x** (capitol letters denote Fourier Transforms of the corresponding lower case variable), and E[ ] is the expected value operator. Note that this spectral definition uses the "boxcar" or "rectangular" window for simplicity. The spectral estimate at each discrete frequency is seen to be a simple algebraic average (i.e, Expected Value) of the "raw" spectra $\tilde{S}$ found from each FFT. The prevalence and computational efficiency of FFT routines makes this the most popular definition of "the spectrum" to most engineers.

But what exactly does the discrete spectrum in Equation 3.1 represent? The first clue to that starts with the strict definition of the Fourier Transform:

$$X(f) \equiv \int_{-\infty}^{+\infty} x(t)\exp(-j\,2\pi f t)\,dt \qquad 3.2$$

This definition requires integration limits of plus and minus infinity. The only way to make this equation numerically-tractable for a non-transient

signal x(t) is to reduce the integration limits to be over a finite time interval, which is the same as zeroing the time series beyond those limits. Thus, a numerically-practical finite FFT is performed on a modified time series x'(t) defined as:

$$x'(t) = \begin{cases} 0 & T \le t \le \infty \\ x(t) & 0 \le t \le T \\ 0 & -\infty \le t \le 0 \end{cases}$$

This is more rigorously accomplished by instead defining a window w(t):

$$w(t) = \begin{cases} 0 & T \le t \le \infty \\ 1 & 0 \le t \le T \\ 0 & -\infty \le t \le 0 \end{cases}$$

such that x'(t)=x(t)w(t) for all t. This definition of w(t) is the well-known boxcar or rectangular window. When this is substituted into Equation 3.2 the non-zero resulting expression becomes:

$$X_w(f) \equiv \int_0^T [x(t)w(t)]\exp(-j\,2\pi f t)\,dt \qquad 3.3$$

where the w subscript denotes the windowed transform. Note that this is still based on continuous time, and yields a transform defined over all frequencies.

The integrand in Equation 3.3 is a product. It is a property of the Fourier Transform that a product in one domain (frequency or time) is a convolution of the transforms in the other domain (time or frequency). Therefore,

$$X_w(f)=\Im\{x(t)w(t)\}=\Im\{x(t)\}\otimes\Im\{w(t)\}$$

$$=X(f)\otimes W(f) \qquad 3.4$$

where the Fourier Transform operation on generic variable z(t) is denoted with the symbol $Z(f)\equiv\Im\{z(t)\}$ and $\otimes$ denotes the convolution operator. For the boxcar window, the Fourier transform W(f) is given by:

$$W(f)=\frac{\sin(\pi f[2T+1])}{\sin(\pi f)} \qquad 3.5$$

The consequence of converting the infinite Fourier transform of Equation 3.2 into the finite (windowed) Fourier transform of Equation 3.3 is that the windowed Fourier transform at any given frequency is the integral from plus to minus infinity of the true transform X(f) convolved by W(f). This smears the true transform by an amount proportional to the bandwidth of W(f), which is approximately $^1/_T$. For example, a single sinusoidal signal (constant amplitude and phase at arbitrary frequency f_0) which has a true transform equal to a delta function at f_0 would yield a windowed Fourier Transform that appears as a series of sidelobes (of width $^1/_T$, repeated over all frequencies with decreasing amplitude as $|f-f_0|$ increases). It is also noted that multiple sinusoids within this $^1/_T$ resolution bandwidth cannot be identified. This is the first explanation for the limiting $^1/_T$ resolution of a FFT-based spectrum.

It is not the intent of this presentation to review the "art" of designing and applying spectral windows such as Hanning, Bartlett, Gaussian, etc., nor to

discuss aliasing effects. Later sections will discuss the role that the choice of FFT-length and corresponding number of ensemble averages play in the spectral resolution and variance of the spectral estimates.

There is a second more intuitive route for understanding the limiting resolution bandwidth available from FFT-based spectra. The theory of equations tells us that it is possible to identify at most 2N parameters from 2N discrete measurements (e.g., 3 points allow a quadratic polynomial consisting of mean, linear, and quadratic coefficients and basis vectors, or 3 Hermite coefficients and the corresponding orthogonal basis vectors, or 3 of any other orthogonal or non orthogonal bases). If a summation of sinusoids is chosen as a basis to represent the measurements, then a set of at most N cosine and sine basis vector frequencies is possible from 2N data points. Thus the model becomes:

$$\sum_{i=1}^{N}\left[a_i \cos(2\pi f_i t_m) + b_i \sin(2\pi f_i t_m)\right] = x_m \qquad m = 1,2\ldots 2N \qquad 3.6$$

The objective in signal processing is to identify the a_i and b_i coefficients such that the summation-of-sinusoids model produces the "best" (such as mean square error) fit to the measurements. Also, as of this time the frequencies in Equation 3.6 are as-yet undefined.

Select least squares as the method for identifying the coefficient column vector $\mathbf{c}=\{\mathbf{a}\ \mathbf{b}\}^T$ where $\mathbf{a}=\{a_1\ a_2\ \ldots\ a_N\}^T$ and $\mathbf{b}=\{b_1\ b_2\ \ldots\ b_N\}^T$ (where ^{T} is the transform operator). Proceed by multiplying Equation 3.6 by each

cosine and sine basis vector and integrating (or equivalently, summing for this discrete representation) over the time spanned by m. This results in 2N equations. Sample equations for coefficients a_l and b_l become:

$$a_l \sum_{i=1}^{N} \sum_{m=1}^{2N} \cos(2\pi f_l t_m) \cos(2\pi f_i t_m) +$$

$$b_l \sum_{i=1}^{N} \sum_{m=1}^{2N} \cos(2\pi f_l t_m) \sin(2\pi f_i t_m) = \sum_{m=1}^{2N} x_m \cos(2\pi f_l t_m) \qquad 3.7a$$

$$m = 1, 2 \ldots 2N, \; l = 1, 2 \ldots N$$

and

$$a_l \sum_{i=1}^{N} \sum_{m=1}^{2N} \sin(2\pi f_l t_m) \cos(2\pi f_i t_m) +$$

$$b_l \sum_{i=1}^{N} \sum_{m=1}^{2N} \sin(2\pi f_l t_m) \sin(2\pi f_i t_m) = \sum_{m=1}^{2N} x_m \sin(2\pi f_l t_m) \qquad 3.7b$$

$$m = 1, 2 \ldots 2N, \; l = 1, 2 \ldots N$$

At this point we can rewrite Equation 3.7 more compactly in linear algebra form showing dimensions:

$$\mathbf{A}_{2N \times 2N} \mathbf{c}_{2N \times 1} = \mathbf{d}_{2N \times 1} \qquad 3.8$$

where bold upper case letters denote a matrix and bold lower case letters denote a column vector. The vector **c** was defined above. Elements in the basis matrix **A** are given by:

$$A_{li} \equiv \begin{cases} \sum_{m=1}^{2N} \cos(2\pi f_l t_m)\cos(2\pi f_i t_m) & l = 1,\ldots N,\ i = 1,\ldots N \\ \sum_{m=1}^{2N} \cos(2\pi f_l t_m)\sin(2\pi f_i t_m) & l = 1,\ldots N,\ i = N+1,\ldots 2N \\ \sum_{m=1}^{2N} \sin(2\pi f_l t_m)\cos(2\pi f_i t_m) & l = N+1,\ldots 2N,\ i = 1,\ldots N \\ \sum_{m=1}^{2N} \sin(2\pi f_l t_m)\sin(2\pi f_i t_m) & l = N+1,\ldots 2N,\ i = N+1,\ldots 2N \end{cases} \qquad 3.9$$

while elements in the data vector **d** are defined as:

$$d_l \equiv \begin{cases} \sum_{m=1}^{2N} x_m \cos(2\pi f_l t_m) & m = 1,\ldots 2N,\ l = 1,\ldots N \\ \sum_{m=1}^{2N} x_m \sin(2\pi f_l t_m) & m = 1,\ldots 2N,\ l = N+1,\ldots 2N \end{cases} \qquad 3.10$$

The next and last necessary step is to complete the definition of the basis vector set by defining the frequencies. There is actually complete freedom in doing so; for example, if the signal was suspected to be comprised of R known frequencies (a good example is the set of analytically-available modal frequencies representing the response of a simply supported beam), then that set of frequencies would be the most logical and appropriate choice to define the basis vectors. Choosing such a particular set of frequencies would define a *parametric* model. While this set would be the most natural engineering choice for defining the basis for this problem, it could in general result in a fully-populated basis matrix **A**. This would at best require maximum computational time to solve the problem (either using Gaussian elimination directly or the normal

form for least squares), and at worst it could introduce ill-conditioning or rank deficiency if any frequencies are too close together (perhaps requiring the pseudo-inverse to solve for the best coefficient vector **c**).

More often the set of frequencies contained in the measurements is not known, nor if there even are discrete frequencies. This is precisely the application for which Fourier Series is best suited. The theory of Fourier Series is a well-developed field, and rightfully so - the method has powerful advantages that are too long to list here. The discussion in this section presents one interpretation of Fourier Series modeling that lays the foundation for development and understanding of the new Harmonic Phase Tracking technique in the next Chapter.

Consider the frequency selection task from the perspective of a numerical analyst whose goal is to minimize the amount of computational resources necessary to fit a model (here, a summation of sinusoids) to data. Indeed, the basis vectors in the Fourier Series model are the most optimal set from this purely *computational* point of view. Specifically, the Fourier frequency set requires the least amount of computations to solve for **c** because it results in a diagonal **A** matrix with the ultimate simplicity that $\mathbf{A}=\alpha\mathbf{I}$ where $\alpha={}^{T}/_{2}$ and **I** is the identity matrix. Since the resulting diagonal matrix has no off-diagonal terms, all the basis vectors are orthogonal, and 2N scalar equations are appropriate to calculate the N a_i and N b_i coefficients. There are many other advantages of this Fourier

orthonormal basis vector set that have been used to great advantage in understanding the behavior of Fourier Series (e.g., Parsevals Theorem) but are not relevant here.

Exactly what values are in this set of Fourier frequencies that yields a diagonal **A** matrix with constant values? It is instructive to work for the moment with the reciprocal of frequency, namely the period (P_i), for this discussion. Fourier defined the period for the first basis vector pair (cosine and sine) equal to the length of the data record: $P_1 \equiv 2N$. The next step was to identify P_2 such that the A_{12} and $A_{1(N+1)}$ matrix terms would be zero. This is accomplished by defining $P_2 = {}^{P_1}/_2$, or equivalently, $f_2 = 2f_1$. Continuing, the complete set of Fourier orthogonal periods and frequencies are defined as integer divisors or multiples of the fundamental: $P_i = {}^{P_1}/_i$ and $f_i = if_1$, respectively.

To a structural engineer, this Fourier set of integer harmonic periods corresponds to the engineering concept of orthogonal mode shapes for a simply supported beam. In this application, the Fourier Series would coincidentally also represent a natural engineering basis for that problem.

With this Fourier Series set of frequencies so defined, we can see that they are spaced at intervals of $f_1 = {}^1/_T$. This is the second interpretation of the implicit resolution of any spectrum based on a FFT: sinusoids spaced closer than ${}^1/_T$ cannot be resolved into uncorrelated spectral ordinates.

But there is a related limitation of Fourier Series that is often not appreciated. The orthogonal nature of the Fourier basis vector set does insure that, for a sinusoid with a frequency f_i that exactly matches a frequency in the set, only the a_i and b_i Fourier coefficients at that one frequency will have non-zero values (assuming proper sampling and no noise). This can be seen by substituting a constant amplitude sinusoid defined by $x_m \equiv a_k\cos(2\pi f_k t_m)+b_k\sin(2\pi f_k t_m)$ in Equation 3.10 and noting that only the d_k (corresponding to a_k) and d_{N+k} (corresponding to b_k) entries in the data vector will be non zero. In some very isolated cases, like modeling the response of a simply supported beam, this exact frequency matching may occur.

But the likelihood of the signal frequency exactly matching one of the Fourier Series frequencies is actually very remote; more often the frequencies in the signal will not exactly match the Fourier Series frequencies. When confronted with this more common situation, it is a common misperception that the energy in a sinusoidal signal is proportionally split solely between the two Fourier basis vectors that bound the signal frequency. This is not true. While most of the energy will be accounted for by the bounding basis vectors, a sizable proportion will be spread to all the basis vectors since they are all correlated to the non-integer signal period. Furthermore, there is false comfort in viewing the *spectrum* and concluding by inspection that the energy spread, while non-negligible, is not important. In contrast, the adjacent *amplitudes* (c_i, $i=k\pm2$) are roughly the same order of magnitude as the bounding Fourier

amplitudes. This is illustrated in Figure 3.1 for two 128-point FFTs (with Hamming and boxcar windows) applied to a unit amplitude sinusoid with a non-integer period of $^{128}/_{10.667}$.

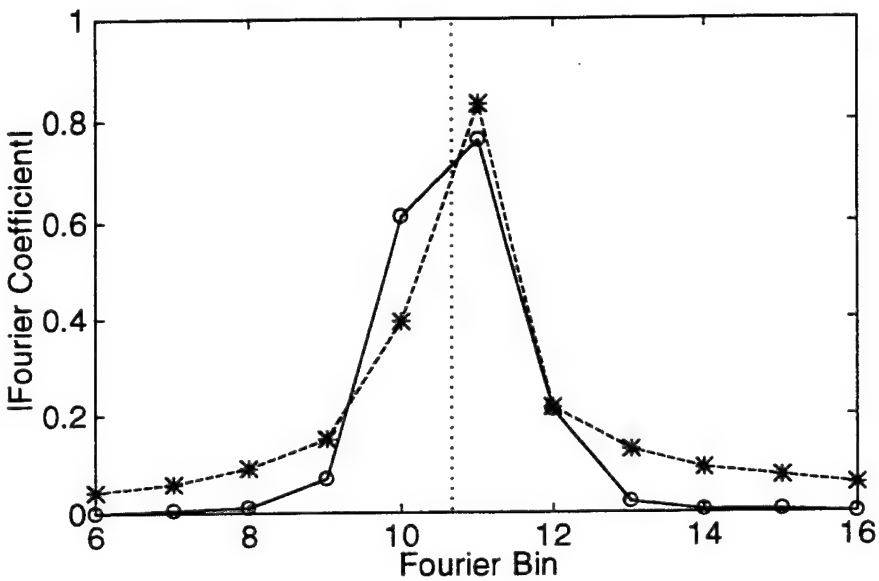

Figure 3.1 Spreading of Fourier Amplitudes with non-integer Harmonic Signal (__o__ = Hamming window; --*-- = boxcar window; .... = exact)

Several insightful observations regarding Fourier Series can be made from Figure 3.1. First, note that the amplitudes are significant for adjacent bins both smaller and larger than the bounding bin numbers of 10 and 11 (note that the kth Fourier frequency can also be labeled as bin number k where bin k corresponds to discrete frequency $^{k}/_{T}$ and period $^{T}/_{k}$; when viewed in this "bin space" k represents the number of cycles of the sinusoid that occur in the FFT interval). This *amplitude* spread is "wider"

than the corresponding spread evident from the amplitude-squared energy in the *spectrum*.

A straightforward explanation for this spread of energy comes from an examination of the Fourier components. Consider the sinusoid used in Figure 3.1 where the non-integer period was not near one of the Fourier harmonics. It is impossible for the transform of this sinusoid to be represented with significant amplitudes only at the two adjacent Fourier bins since the sum of these two component sinusoids would necessarily be a beating signal with a non-constant amplitude. This is illustrated in Figure 3.2 for the same signal used in Figure 3.1. The top subfigure is the original sinusoid; the middle subfigure is the [beating] signal formed by the inverse transform using only the two Fourier harmonics at the adjacent bins; and the lower subfigure is the difference (remainder) between the first two signals (all subfigures use the same scale). Since Parseval's theorem requires energy proportionality between the time and frequency domain representations, the size of the remainder time domain signal does result in significant frequency domain energy (at harmonics other than the adjacent two Fourier bins). Note also that application of a spectral window (Hamming, etc.) would act to suppress the ends of this remainder time series component, thereby reducing much of the variance in both the time and frequency domains.

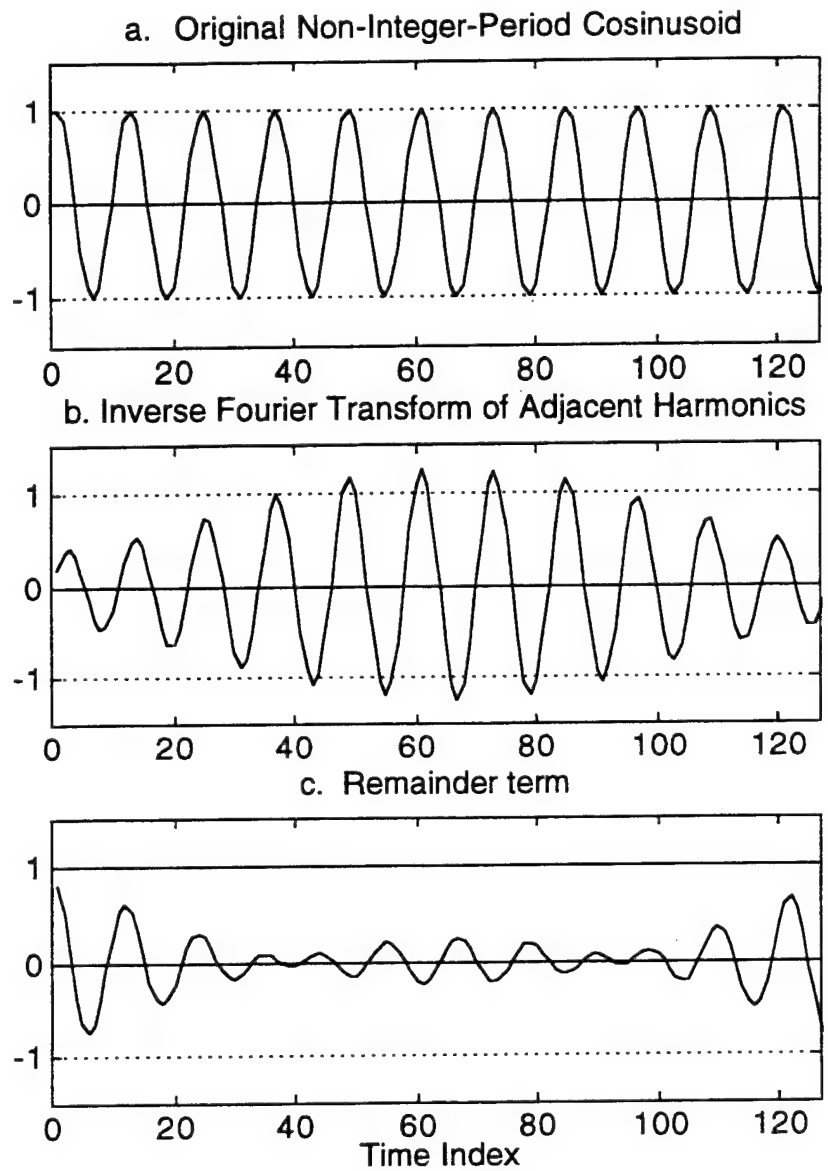

Figure 3.2 Example Decomposition of Inverse Fourier Transform of Non-Integer-Period Sinusoid Used in Figure 3.1

Second, note in Figure 3.1 that the Hamming window does result in significantly lower amplitudes *away* from the major bandwidth, but it is effectively wider than the boxcar window *inside* the bandwidth between bins 10 and 11 (which illustrates the fact that each window has its advantages and disadvantages). Note also that this example clearly illustrates the effect of a window taper in the time domain, which artificially suppresses much of the residual signal variance, thereby reducing the leakage according to Parseval's Theorem. What is not so readily observed in Figure 3.1 is that the spread of the Fourier Series actually produced aliasing (caused by the discontinuity between the adjacent periodic sequences then the folding at the Nyquist frequency) that further complicates interpretation of any windowed transform as well as any subsequent spectrum.

Let us also recognize that such a sinusoidal signal with a frequency that is a "non-integer" fraction of the FFT length can be analytically described using just three parameters (or degrees of freedom): cosine amplitude, sine amplitude, and frequency. However, because of the spread inherent in the Fourier coefficients it will require all of the 2N Fourier series coefficients to model it; while the basis matrix is diagonal and hence "efficient" in terms of computational resources, the right hand side data vector **d** has all non-zero entries, leading to all non-zero coefficients in **a** and **b**. This is hardly an "efficient" mapping of such a simple signal.

Finally, it is instructive to examine confidence limits for spectral ordinates. It is conventionally assumed that since the equivalent signal phases are arbitrary and independent at each Fourier frequency, then the corresponding "raw" a_i and b_i coefficients are Gaussian distributed for each independently analyzed FFT realization of the time history. Second, it is known that a Gaussian input to a linear system produces a Gaussian output. Since the Fourier transform is a linear operator, then the real and imaginary transforms are likewise Gaussian distributed. The spectral estimates are then the sum of two squared Gaussian variables ($S_i=a_i^2+b_i^2$), which is described by a Chi-squared distribution with 2 (for the double sum) degrees of freedom (DOF). Unfortunately, a Chi-squared variable with 2 DOF has large uncertainty, so a raw spectral estimate based on one transform is quantitatively poor. This explains the need for the use of the Expected Value Operator in Equation 3.1 (i.e., it is not consistent to assume that increasing the length of the FFT places more component cycles into the FFT integration and that such a transform is therefore statistically more accurate than a transform based on a shorter FFT). This explains in mathematical terms the need for ensemble averaging to confidently describe a spectrum. It is well-known that the normalized uncertainty for spectral estimates is given the Chi-squared distribution and the number of independent "raw" transforms (N):

$$\varepsilon(\hat{S}_i / S_i) \equiv \sqrt{1/N} \qquad i = 0,1,\ldots N/2 \qquad 3.11$$

For example, 100 ensemble averages would yield reasonable ±10 percent uncertainty limits for the spectrum, while 1 (no average) would yield 100 percent uncertainty limits. Collecting this much wave data is sometimes possible in a laboratory setting (subject to reflections and other "tank buildup" phenomena). As discussed in the next Chapter, to achieve this accuracy an ocean wave field would have to be stationary for many hours, which is most often not the case.

Now we can return to Equation 3.1 and summarize what to expect when it or any of the other traditional methods is used to produce a spectrum from a properly sampled wave or similar time history:

- a set of windowed FFTs will generally resolve the distribution of variance versus frequency in a statistically averaged sense.
- however, a large number of independent transform estimates are required to yield reasonable error bounds for that variance distribution, requiring long-term stationarity in the signal.
- if discrete sinusoids are present, they cannot be resolved if closer than $^1/_T$ in frequency where T is the length of the FFT.
- sinusoids with non-integer multiples of T produce significant smearing ("leakage") and is difficult to identify and resolve.
- this smearing also makes identification of nonstationarities (like changing amplitude, phase, or frequency) difficult.
- the FFT is a non optimum mapping (3 to 2N parameters) for a simple sinusoid when the FFT segment length divided by the signal period is not an integer.

Only now, having completed this review of traditional spectral analysis, can a summary quote from Kay (1988, p6) be fully appreciated:

> *Spectral estimation is a* preliminary data analysis tool. *A spectral estimate should not be used to answer* specific *questions about the data, such as whether a resonance is present, but only to suggest possible hypotheses.*

This conclusion also applies to the study of ocean waves and is reaffirmed by a quote from Goda (1985, p213):

> *Of course, we are always free to analyze an irregular time-varying function, ... in the form of a Fourier series without attaching any particular physical meaning.*

The purpose of this section has been to interpret Fourier Series from one point of view and illustrate some of its weaknesses as they apply to the study of ocean waves in the next section and the next Chapter. This has been done with the full realization and appreciation that for general signal analysis applications the overall strengths of Fourier analysis far outweigh these weaknesses.

3.3 Subspace Estimation Techniques

The techniques in the previous section can be applied regardless of the signal characteristics as long as it is stationary. The penalty for that universality is a loss of resolution and any "local" behavior as a result of the ensemble averaging. In this section an entirely different family of techniques is described. These techniques are classified as "high resolution" parameter estimation (rather than spectral estimation) techniques because they assume a signal model defined as a small and known number of sinusoids in the presence of white noise, and they proceed to estimate that precise set of unknown parameters. This is quite restrictive compared to the spectral methods, but when applied properly these subspace techniques are very powerful.

In a sense, the foundation for all of these methods is the following key theorem (Roy, 1987):

Theorem 3.1 (Wold's Predictive Decomposition - Wold, 1938). A (weakly) stationary process is decomposable into two (weakly) stationary processes orthogonal to each other, $s(t) \perp n(t)$:

$$x(t)=s(t)+n(t)$$

such that

$$n(t) = \sum_{i=0}^{\infty} c(t-i)\xi(i)$$

where c denotes coefficients in a moving average description of n(t), the innovations ξ are orthonormal, and s(t) is deterministic while n(t) is completely non deterministic.

This theorem says that it is always possible to decompose a stationary realization into the sum of a deterministic (s) and a non deterministic (n) component, and that the two spaces will be orthogonal. The theorem has been extended since then to allow for components of the non deterministic space in the deterministic space (trading subspaces) and visa versa. (This concept has more recently been made practical to apply by the availability of singular value decomposition, as will be shown later in this Chapter.) The need for such extensions is seen, for example, in interpreting a Fourier Spectrum (i.e., Series) and arbitrarily defining some components as "the signal" with the rest (including some small and perhaps overlooked deterministic signal components) automatically defined to be "noise" components. All of the subspace techniques manipulate these two spaces in various ways.

One way to understand these spaces it to examine a deterministic signal comprised of one sinusoid with constant (unit) amplitude and frequency and random phase. The (auto)covariance vector versus lag for this signal is defined as:

$$c_{xx}(\tau) \equiv E[x(t)x(t+\tau)] \qquad 3.12$$

where E[ ] is the expected value operator. Substitute a unit amplitude sinusoid $\mathbf{x}(t)\equiv\cos(2\pi f_0 t+\phi)$ with arbitrary phase to find:

$$\mathbf{c}_{xx}(\tau)=\cos(2\pi f_0\tau) \qquad 3.13$$

for all τ. For a discrete time signal Equation 3.13 becomes simply

$$\mathbf{c}_{xx}(\tau_m)=\cos(2\pi f_0\tau_m) \qquad 3.14$$

for $\tau_m = m\Delta t$ and m=1,2,....M. Now, construct a [true] covariance matrix by assembling columns defined as cyclically shifted $\mathbf{c}_{xx}$:

$$\mathbf{C}_{xx} \equiv \begin{bmatrix} c_{xx}(0) & c_{xx}(-1) & c_{xx}(-2) & \cdots & c_{xx}(-M) \\ c_{xx}(1) & c_{xx}(0) & c_{xx}(-1) & \cdots & c_{xx}(-M+1) \\ c_{xx}(2) & c_{xx}(1) & c_{xx}(0) & \cdots & c_{xx}(-M+2) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ c_{xx}(M) & c_{xx}(M-1) & c_{xx}(M-2) & \cdots & c_{xx}(0) \end{bmatrix} \qquad 3.15$$

For a harmonic signal without noise $c_{xx}(-\tau)=c_{xx}(\tau)$ and the matrix is symmetric (for infinite available data). In this idealized case, since this matrix uses just one column basis vector to construct every other column by simple shifting, it is a rank 1 matrix. If a standard singular value decomposition (e.g., Leon, 1994) of $\mathbf{C}_{xx}$ was done defined by

$$\mathbf{U}\Sigma\mathbf{V}^T\equiv\mathrm{svd}(\mathbf{C}_{xx}) \qquad 3.16$$

results (and simple inspection of the $\mathbf{C}_{xx}$ matrix) show:

- one non zero singular value $\sigma_1=\Sigma_{11}$,
- the first left singular vector $\mathbf{u}_1=\mathbf{c}_{xx}$ (that is, column 1), and
- the first right singular vector $\mathbf{v}_1=\mathbf{c}_{xx}^T$ (that is, the top row).

Thus, in this case,

$$\mathbf{C}_{xx}(t)=\sigma_1 \mathbf{u}_1 \mathbf{v}_1^T \qquad 3.17$$

The fact that the rank of this covariance matrix is independent of the size of the matrix $\mathbf{C}_{xx}$ is the most important fact exploited by subspace methodologies. As this first simple example illustrated, finding the rank of $\mathbf{C}_{xx}$ was equivalent to finding the number of components in the signal.

Now, add two real world complications to this idealized example. First, add noise to the unit amplitude sinusoidal signal so that $\mathbf{x}(t)\equiv s(t)+n(t)$ where $s(t)=\cos(2\pi f_0 t+\phi)$. Second, assume a finite length of measurements of $x(t)$. These conditions affect $\mathbf{C}_{xx}$ in the following way:

1. For this special case (even assuming infinite data length), $\hat{\mathbf{C}}_{xx}$ is an estimate of the true covariance matrix $\mathbf{C}_{xx}$ of the signal summed with the covariance matrix $\mathbf{C}_{nn}$ for the noise:

$$\hat{\mathbf{C}}_{xx}=E[x(t)x(t+\tau)]=E[(s(t)+n(t))(s(t+\tau)+n(t+\tau))]$$

$$=E[s(t)s(t+\tau)]+E[s(t)n(t+\tau)]+E[s(t+\tau)n(t)]+E[n(t)n(t+\tau)]$$

$$=E[s(t)s(t+\tau)]+E[n(t)n(t+\tau)]$$

$$\hat{\mathbf{C}}_{xx}=\mathbf{C}_{xx}+\mathbf{C}_{nn} \qquad 3.18$$

Note that if the signal consisted of two sinusoids $s_1(t)$ and $s_2(t)$ at arbitrary frequencies, then the covariance would not be a simple sum of the two component covariance functions because if the

sinusoids are not orthogonal then the cross terms would contribute. This complicates the covariance function for multi-component signals.

2. Only the estimate $\hat{\mathbf{C}}_{xx}$ of $\mathbf{C}_{xx}$ can be constructed from the data because of the stochastic noise matrix and the fact that the true covariance matrix cannot be found due to the finite length of data.

The subspace methods were first developed for array directional resolution applications. In these problems the signal is known to be a sinusoid and assumed to be planar so that any phase difference between sensors is due solely to their spatial separation. For these problems the previous covariance matrix is approximated by a "sample" covariance matrix defined directly from the measurements by ensemble averaging "raw" matrices:

$$\hat{\mathbf{C}}_{xx} \equiv \frac{1}{N} \sum_{i=1}^{N} \mathbf{E}\left[\mathbf{x}_i \mathbf{x}_i^T\right] \qquad 3.19$$

where $\mathbf{x}_i$ is a zero-mean, finite length vector from the measurement vector $\mathbf{x}$ with the same length as needed for the covariance matrix. To illustrate this, return to the previous single sinusoid deterministic example and replace the covariance matrix with this sample covariance matrix. Now, each column and row has two basis vectors - one related to $\alpha\cos(2\pi f_0 \tau_m)$ and another related to $\beta\sin(2\pi f_0 \tau_m)$, where α and β are functions of the phase angle ϕ.

Without added noise, a singular value decomposition of this illustrative sample covariance would return two non-zero singular values, and two left and two right singular vectors. Thus, the sample covariance matrix serves the same purpose as the exact covariance matrix but with double the rank.

With any amount of added noise, the rank of the sample covariance matrix becomes full because the noise covariance in Equation 3.19 is full rank. For this more typical case, the modified procedure starts with inspection of the singular values. In many cases, the singular values (from the diagonal of matrix Σ) will all be approximately the same (at least in order of magnitude) after the first r values:

$$\sigma_{kk} \approx \sigma_{NN}, \qquad 2r+1<k<N \tag{3.20}$$

These are now arbitrarily defined to be noise contributions. Thus, the rank of the matrix is again concluded to be 2r, which in turn defines r component sinusoids. Of course, this is a non unique definition if there is continual rather than an obvious demarcation in the set of singular values.

Having determined the number of components assumed to be in the signal, most of the methods then define one or both of the signal and noise subspaces (**H** and **E**, respectively) using left singular (column) vectors:

$$\mathbf{H} = \begin{bmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \cdots & \mathbf{u}_{2r} \end{bmatrix} \tag{3.21}$$

and

$$\mathbf{E} = \begin{bmatrix} \mathbf{u}_{2r+1} & \mathbf{u}_{2r+2} & \cdots & \mathbf{u}_N \end{bmatrix} \qquad 3.22$$

It is a property of the singular value decomposition that these spaces are orthogonal. But also note that neither **H** nor **E** are unique, even when the noise is very small and does not significantly interfere with the predominant singular vectors. For example, for a signal consisting of two sinusoids, the left singular vectors in **H** would look like a pair of in- and out-of-phase modulated (beating) basis vectors rather than the "expected" ($\mathbf{H}_e$) space consisting of two cosines and two sines; this is because these two left (with the two right) singular vectors define the best rank-two approximation to the original covariance matrix (see Equation 3.17). This is illustrated in Figure 3.3. Although **H** and $\mathbf{H}_e$ are different, they are linearly dependent (as seen by the direct algebraic trigonometric relationship between the vectors in **H** and $\mathbf{H}_e$ in this two sinusoid example). Therefore both spaces are orthogonal to **E**.

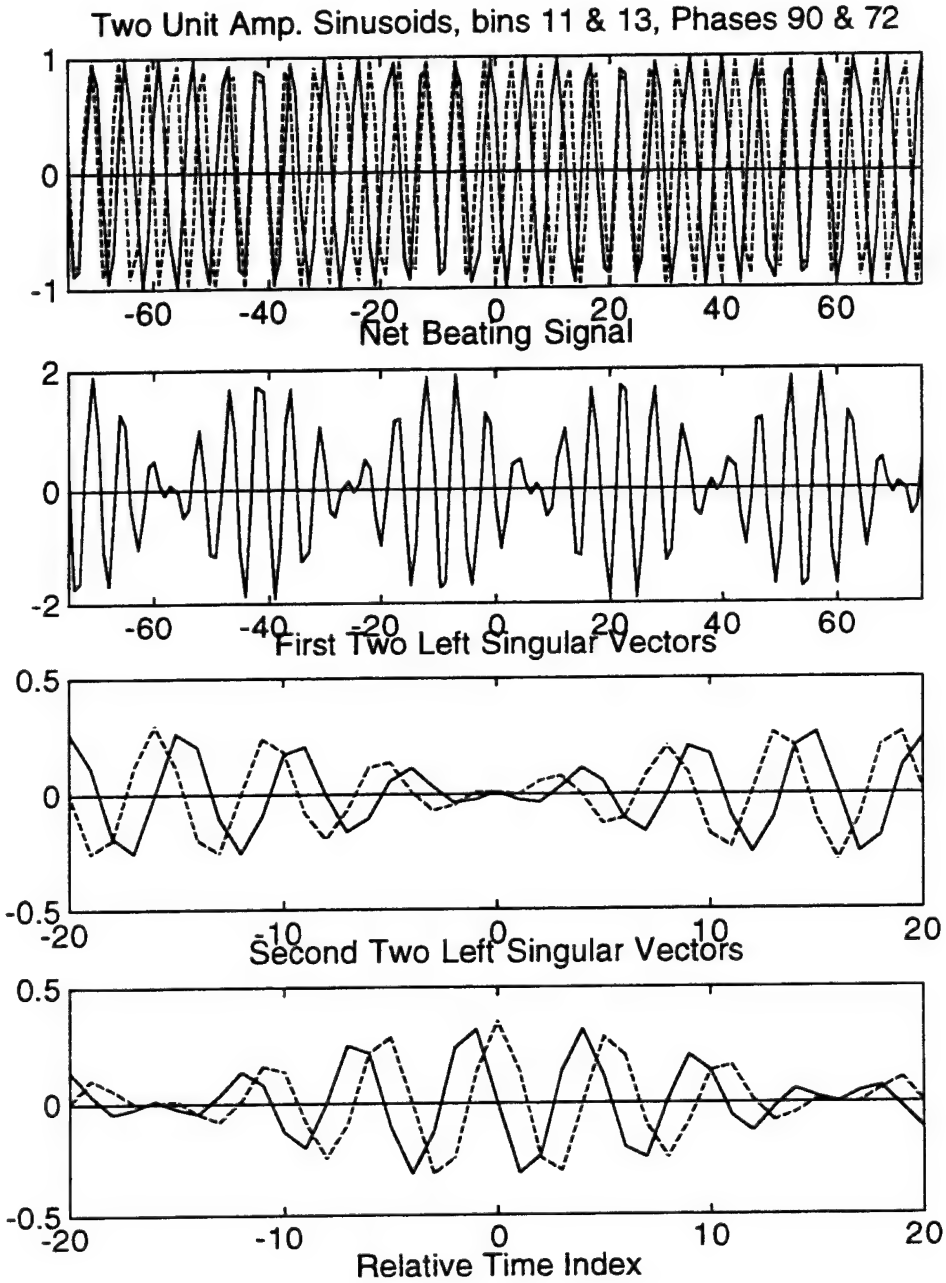

Figure 3.3 Example of Singular Value Vectors for a Deterministic Signal Comprised of Two Sinusoids

The next task for the subspace methods is to estimate the signal frequencies. This is done in a variety of very different ways among the methods.

For example, there are straightforward search techniques such as with the MUSIC algorithm, which was first developed in 1979 and is still considered best for general applications (Schmidt, 1979; see also Roy, 1987 and Naidu, 1996). This technique is based on the fact that a sinusoid with a trial frequency that matches one of the true signal frequencies will be in the signal space, and therefore by definition it will be orthogonal to the entire noise space. MUSIC assumes that for a sinusoidal vector at trial frequency f_k described by either

$$e_k = \left[1 \ \cos(2\pi f_k \Delta t) \cdots \cos(2\pi f_k (L-1)\Delta t)\right] \qquad 3.23a$$

or

$$e_k = \left[0 \ \sin(2\pi f_k \Delta t) \cdots \sin(2\pi f_k (L-1)\Delta t)\right] \qquad 3.23b$$

that if f_k is equal to a true frequency then the subspace orthogonality condition requires that

$$\mathbf{E}^T \mathbf{e}_k = \mathbf{0} \qquad 3.24$$

If noise is present such that only estimates of **H** and **E** are available ($\hat{\mathbf{H}}$ and $\hat{\mathbf{E}}$, respectively), then a modified criterion is needed:

$$\mathbf{e}_k^T \hat{\mathbf{E}} \hat{\mathbf{E}}^T \mathbf{e}_k = \text{ local minimum} \qquad 3.25$$

or, as usually implemented, define a reciprocal function and equivalently search for maxima of the "null" function

$$\hat{S}(f_k) \equiv \frac{1}{\mathbf{e}_k^T \hat{\mathbf{E}} \hat{\mathbf{E}}^T \mathbf{e}_k} \qquad 3.26a$$

or

$$\hat{S}(f_k) \equiv \frac{1}{\left\| \hat{\mathbf{E}}^T \mathbf{e}_k \right\|_2} \qquad 3.26b$$

where $\| \ \|_2$ is the 2-norm of the vector. This is loosely referred to as a "spectrum" even though it does not represent the distribution of variance with frequency. A fine mesh of trial sinusoidal $\mathbf{e}_k$ vectors are generated and used to calculate $\hat{S}$ (this function will have only so much inherent resolution but it can be calculated at any desired frequency interval). By inspection, the largest r peaks of $\hat{S}$ are the best estimates of the true frequencies present in the underlying signal.

Other subspace methods include: Capon, Prony, Pisarenko, and ESPRIT, with many different implementations for each method (motivated by different assumptions about the noise characteristics or how well the number of components in the signal is known). The subspace rotation approach used in the ESPRIT techniques is particularly interesting (Roy, 1987). ESPRIT offers significant advantages over MUSIC for some applications.

Before summarizing the practical issues regarding the use and interpretation of these subspace methods, it would be useful to contrast their frequency resolution to the "traditional" spectral analysis

techniques. Unfortunately, there are no exact analytical expressions for the variance of the frequency estimates from these methods (Kay, 1988). But for our purposes they are not necessary, since these methods show dramatically better resolution regarding frequency estimation than the traditional techniques.

With such dramatically-improved performance in resolution, why wouldn't one of these subspace methods like MUSIC or ESPRIT always be used in signal processing? Do they have assumptions about the signal and/or noise that limit their general applicability? (Note that this question ties back directly to ocean wave issues raised back in Chapter 2.)

To start with, these methods universally assume that the signal is known to be a summation of a [small] number of sinusoids. Thus, it can be immediately stated that it is not appropriate to apply them to any other physical situation.

Second, the number of sinusoids must be "small". While this term is never quantified, none of the dissertations or texts referenced in this review ever considered more than a half dozen at most. This is probably due to the fact that the main application for these techniques is estimating the best direction-of-arrival of radar signal on a linear array, and the number of radar sources (targets) is typically small in those applications. The use of these methods for other situations where there might be three or more

dozen sinusoidal signals (say, in the general case of an ocean wave field) is not strictly beyond their capabilities in the presence of no noise, but it is without precedent. The use of these techniques for noisy signals with a large number of sinusoids is questionable.

Knowledge of the spectral distribution of the noise is crucial to many of the methods that rely on finding eigenvalues of the sample covariance matrix (such as Pisarenko decomposition). Consider the case with a known signal and noise: if the true covariance matrix $\mathbf{C}_{xx}$ given in Equation 3.15 and the $\mathbf{e}_k$ vectors from Equation 3.23 are defined using the exact frequencies and known number of signals (r), then

$$\mathbf{C}_{xx} = \sum_{k=1}^{r} \mathbf{P}_k \mathbf{e}_k \mathbf{e}_k^H \qquad 3.27$$

where $\mathbf{P}$ is the power for each signal and ^{H} is the Hermitian operator (necessary if the complex definition of $\mathbf{e}_k=\{\exp(j2\pi ft)\}$ is used). An eigendecomposition of $\mathbf{C}_{xx}$ in terms of eigenvalues λ_i and basis vectors ξ_i would yield

$$\mathbf{C}_{xx} = \sum_{k=1}^{r} \lambda_k \xi_k \xi_k^H \qquad 3.28$$

Note that the rank is still r. When uncorrelated additive noise is present, it can be decomposed similarly:

$$\mathbf{C}_{nn} = \sum_{k=1}^{N} \Lambda_k \Xi_k \Xi_k^H \qquad 3.28$$

But here, the rank is N, equal to the length of the data vector. Using Equation 3.18, one analytical form for the estimated covariance matrix would be the sum of two matrices with different ranks

$$\hat{C}_{xx} = \sum_{k=1}^{r} \lambda_k \xi_k \xi_k^H + \sum_{k=1}^{N} \Lambda_k \Xi_k \Xi_k^H \qquad 3.29$$

If an eigendecomposition was done directly on this estimated matrix $\hat{C}_{xx}$, it, too, would yield a second analytical form consisting of a full rank summation of eigenvalues and basis vectors

$$\hat{C}_{xx} = \sum_{k=1}^{N} \chi_k v_k v_k^H \qquad 3.30$$

Once the rank r of the signal subspace is known or estimated, Equation 3.30 is separated into two summations:

$$\hat{C}_{xx} = \sum_{k=1}^{r} \chi_k v_k v_k^H + \sum_{k=r+1}^{N} \chi_k v_k v_k^H \qquad 3.31$$

Many of the subspace techniques need to relate Equations 3.29 and 3.31 to estimate the signal eigenvalues χ_k, k=1,....r, and in turn the signal frequencies. That is not possible in the present form. Two additional steps are required. The first uses the known fact that since the two signal matrices defined by the first summations in Equations 3.29 and 3.31 both span the same (signal) subspace, the matrices are interchangeable. The second step is to take advantage of the fact that any orthonormal matrix times its Hermitian yields the identity matrix. This can be incorporated here - but only if the noise is assumed to be white noise such that the

noise basis vectors can be moved outside the summation. The noise covariance matrix can then be simplified to

$$\begin{aligned} C_{nn} &= \sigma^2 \sum_{k=1}^{N} \Xi_k \Xi_k^H \\ &= \sigma^2 \sum_{k=1}^{N} v_k v_k^H \end{aligned} \qquad 3.32$$

Substitute both of these changes into Equation 3.31 to yield

$$\hat{C}_{xx} = \sum_{k=1}^{r} (\chi_k + \sigma^2) v_k v_k^H + \sigma^2 \sum_{k=r+1}^{N} v_k v_k^H \qquad 3.33$$

Comparing the true and estimated signal eigenvalues in Equations 3.28 and Equation 3.33 shows that the estimated eigenvalues λ_k are larger than the true eigenvalues χ_k by σ^2

$$\lambda_k = \chi_k + \sigma^2, \qquad k=1, \ldots. r \qquad 3.34$$

Of course, Equation 3.34 can then be used to improve the estimated eigenvalues, but this correction is only as good as the white noise assumption used to generate this Equation. Generally speaking, all of the subspace methods require either *a priori* knowledge of the noise covariance matrix or are forced to assume that it is uncorrelated white noise. This is not always appropriate, so it can introduce error into estimates made using these techniques.

But what about other subspace techniques that do not rely on eigenvalue estimates? One common problem is the need to *a priori* estimate the rank of the signal subspace, and its effect on the results varies among the methods. This is straightforward if there is a sharp demarcation in the

singular values, but that is often not the case. There has been a great deal of attention in the signal processing community directed towards quantifying techniques to simply find the signal subspace rank, but none are yet considered "definitive."

Some of the techniques require constant amplitude sinusoids, and, for array direction finding applications, arrays with known and constant spacing. Other methods require that the sample covariance matrix be Toeplitz, which is typically violated when using finite lengths of data (this requirement spawned sub methods to force symmetry or use singular value decomposition or other techniques to improve the covariance matrix). The powerful ESPRIT technique assumes low rank, planar, independent narrowband signals of known center frequency, which can be restrictive.

In summary, the new subspace techniques have clearly demonstrated higher resolution capabilities as compared to traditional spectral analysis techniques. Many of them rely directly on the orthogonality of assumed signal and noise subspaces, which is itself subjective, while others continue on with additional mappings (such as the eigenvalue estimation concept that make it impossible to analytically quantify the uncertainties). But the price for the improved resolution from these techniques is high in terms of the restriction that the signal consist of a small number of sinusoids, usually with uncorrelated and additive white noise.

3.4 Wavelet and Other Local Techniques

Wavelets offer a useful alternative for some signal processing applications so their properties are briefly described in this section. Three other studies of "local" ocean wave properties using alternative techniques are reviewed also.

Before considering wavelet analysis, it is instructive to summarize Fourier Series analysis as modeling with a superposition of constant parameter orthogonal sinusoids. The essential difference between Fourier and wavelet analysis is that wavelets incorporate a time- and frequency-dependent envelope (like a spectral window) to *each* of the basis vectors (which are not typically sinusoids), whereas [unwindowed] Fourier basis vectors are constant amplitude throughout the segment being analyzed. These wavelet envelopes vary with frequency to retain a constant "volume" of each basis vector in the time-frequency plane. To aid in visualizing wavelets, imagine a "characteristic" wavelet defined as a constant parameter sinusoid modulated by a vary narrow Gaussian window in time such that their product yields only a half dozen cycles of the sinusoid before the amplitude is effectively zero. Now maintain this "half dozen cycles" for each sinusoidal basis vector; low frequency sinusoids will have relatively wide envelopes while high frequency sinusoids will have very narrow envelopes. Typically, the "constant volume" requirement also results in wavelet amplitudes that are likewise proportional to the frequency. In application these characteristic

wavelets are centered at a time step of interest and then fitted to the signal just like with Fourier Series to yield a set of "participation factors". The center time is shifted by a small amount and the analysis is repeated. Ultimately, these factors are displayed versus time and frequency to yield a "time-frequency" distribution. There are many implementations of wavelet basis vectors and modulating envelopes. Wang (1995) presents a very readable review of wavelet analysis; see also Chu (1996) and Donelan, et al (1996) for applications to ocean waves.

Note that the wavelet basis vectors are only non zero for a very limited number of cycles (a half dozen in this example). This makes wavelet analysis ideally suited for signals that exhibit sharp discontinuities; indeed, wavelets are intended primarily for detection of this and similar events. But a less ideal characteristic of wavelets is that they exhibit increasing passband, which means that the frequency resolution is inversely proportional to the frequency (Naidu, 1996). But the orthogonal basis vector set, coupled with the increasing resolution bandwidth due to the time-frequency ambiguity, makes the application of wavelets suspect for understanding ocean wave fields.

At least two studies have been reported that used instantaneous frequency (IF) to describe "local" wave behavior (Huang, et al, 1992; Gran and Bitner-Gregersen, 1983). As described in more detail in Chapter 4 and Appendix C, the algebraic interpretation of the IF for multicomponent signals is difficult; in addition, Huang, et al used arguments based on a

fractal analysis of ocean waves to question whether ocean waves are differentiable, with the consequence that the IF may be undefined. (This issue will be addressed in Chapter 5.) Regardless, these analyses only yield "effective" scalar local descriptors of the underlying ocean wave process in terms of a time-varying instantaneous envelope and time-varying instantaneous frequency - neither of which provide much information regarding the actual physics of the component waves.

Borgman, et al, (1993) present a new method for estimating the evolution of the spectrum for nonstationary processes such as ocean waves. This proposed method uses the usual set of orthogonal Fourier Series frequencies, but assumes that the component amplitudes over each segment vary according to a quadratic (or cubic) polynomial. Allowing for amplitude variations does improve some of the leakage inherent in FFT processing, but because this technique still uses the fixed Fourier Series frequency set it still has sizable leakage problems, particularly with the phases.

3.5 Chapter Summary

As outlined in this Chapter, there are presently two general categories of signal processing techniques that can be even considered for identifying the "local" behavior of ocean waves - low resolution and high resolution techniques. Both suffer from significant limitations (time-frequency resolution, low rank, etc.) and the need for subjective decisions (such as defining the rank of the signal or noise subspace). Similar arguments apply to wavelet analysis.

A final quote from Goda (1985, p212) nicely summarizes the value of the present body of signal processing theory to the understanding of the physics of ocean waves:

> *"The interpretation of random sea waves as a linear superposition of free progressive waves is an assumption, the correctness of which cannot be proven but, rather, must be supported through evidence of agreement between the properties of real sea waves and those derived from the mathematical model. . . . . Thus, the component waves [from an infinite summation of infinitesimal amplitudes] do not represent physical reality in themselves.*

Thus, even with all of the recent advances, there is still a clear signal processing deficiency that has made it impossible to satisfactorily identify and describe the local behavior of ocean waves.

[blank]

CHAPTER 4

DESCRIPTION OF HARMONIC PHASE TRACKING METHOD

4.1 Chapter Overview.

The information presented in the previous two chapters established that there is no existing technique capable of adequately describing an ocean wave field. Chapter 3 classified stochastic signal processing techniques as either spectral analysis (low resolution) or parameter estimation (high resolution). Spectral analysis is not an optimum choice for wave analysis because the time scale for stationarity is more on the order of tens of minutes rather than the hours needed for spectral ensemble averaging. On the other hand, the newer parameter estimation techniques have the potential of returning much more information and resolution than the spectral techniques, but for this application their use would be difficult because the noise characteristics are not known (neither the covariance matrix nor the standard deviation), and the rank (number of sinusoidal components) can be rather high.

The new Harmonic Phase Tracking (HPT) technique presented in this chapter combines the best features of both the low and high resolution descriptors. For example, the HPT technique does use the same *a priori* assumption as the high resolution techniques, namely, that the signal is a finite summation of constant parameter sinusoids with unknown and arbitrary frequencies, amplitudes and phases. However, it is superior to existing high resolution techniques because it makes no *a priori* assumptions regarding the noise and can accomodate a large number of sinusoids. This chapter presents the theoretical basis for the new technique by examining some constant parameter signals, starting with a single sinusoid with and without noise (Sections 4.2 to 4.5), then a multiharmonic signal with and without noise in Section 4.6. Section 4.7 describes the methodology for identifying the initial vector (including rank) of estimated frequencies in the signal. The essence of this proposed technique is first identifying the true signal frequencies as accurately as possible, then subsequently using that information for the estimation of the amplitude and phase vectors.

4.2 Algebraic Development of HPT as Applied to a Single Sinusoid

The fundamentals of the new technique will be illustrated in this and the next two sections using constant parameter signals. Consider first a continuous unit amplitude signal defined by:

$$x(t) \equiv \cos(2\pi f t) \qquad 4.1$$

The frequency is unknown, and a zero phase is used for simplicity.

The objective is to estimate the frequency, then the amplitude and phase of this signal. Define the estimated signal (i.e, the model) as:

$$\begin{aligned} \hat{x}(t) &\equiv \hat{c}\cos(2\pi \hat{f} t + \hat{\theta}) \\ &\equiv \hat{a}\cos(2\pi \hat{f} t) + \hat{b}\sin(2\pi \hat{f} t) \end{aligned} \qquad 4.2$$

where the "^" symbol denotes an estimate of the true variable. The estimated (not the true) frequency is assumed known. Estimates for the true amplitude $\hat{c}$ and phase $\hat{\theta}$ are both based on finding the component amplitudes $\hat{a}$ and $\hat{b}$. For the purposes of this section it is not important to justify how the frequency was estimated; that is the topic of Section 4.7.

Apply the usual least squares approach to find $\hat{a}$ and $\hat{b}$. Define an error function as the integral of the squared error over some time integration interval:

$$Q = \int_{-\xi}^{\xi} [\hat{x}(t) - x(t)]^2 \, dt$$

$$= \int_{-\xi}^{\xi} [\hat{a}\cos(2\pi\hat{f}t) + \hat{b}\sin(2\pi\hat{f}t) - \cos(2\pi f t)]^2 \, dt \qquad 4.3$$

Take partial derivatives with respect to $\hat{a}$ and $\hat{b}$; set each derivative equation equal to zero corresponding to the optimum values; and rearrange to yield:

$$\hat{a}\int_{-\xi}^{\xi} \cos^2(2\pi\hat{f}t)\, dt + \hat{b}\int_{-\xi}^{\xi} \cos(2\pi\hat{f}t)\sin(2\pi\hat{f}t)\, dt =$$

$$\int_{-\xi}^{\xi} \cos(2\pi f t)\cos(2\pi\hat{f}t)\, dt$$

$$\hat{a}\int_{-\xi}^{\xi} \cos(2\pi\hat{f}t)\sin(2\pi\hat{f}t)\, dt + \hat{b}\int_{-\xi}^{\xi} \sin^2(2\pi\hat{f}t)\, dt = \qquad 4.4$$

$$\int_{-\xi}^{\xi} \cos(2\pi f t)\sin(2\pi\hat{f}t)\, dt$$

Note the symmetrical limits, requiring that time is a zero-mean vector in Equation 4.4. Evaluate the trigonometric analytical integrals:

$$\int_{-\xi}^{\xi} \cos^2(2\pi\hat{f}t)\, dt = \xi(1 + \mathcal{S}_{4\pi\hat{f}\xi}) \qquad 4.5$$

where

$$\mathcal{S}_{\alpha} \equiv \frac{\sin\alpha}{\alpha} \qquad 4.6$$

is the well-known sinc function and is used throughout this Chapter and Appendix C. Continuing with terms on the left hand side of 4.4,

$$\int_{-\xi}^{\xi} \cos(2\pi\hat{f}t)\sin(2\pi\hat{f}t)\,dt = 0$$

$$\int_{-\xi}^{\xi} \sin^2(2\pi\hat{f}t)\,dt = \xi(1 - \mathcal{S}_{4\pi\hat{f}\xi}) \qquad 4.7$$

While the integrals on the right hand side become:

$$\int_{-\xi}^{\xi} \cos(2\pi\hat{f}t)\cos(2\pi f t)dt = \xi(\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi})$$

$$\int_{-\xi}^{\xi} \sin(2\pi\hat{f}t)\cos(2\pi f t)\,dt = 0 \qquad 4.8$$

where $f_\Delta \equiv \left|\hat{f} - f\right|$ and $f_\Sigma \equiv \hat{f} + f$ are the unknown difference and sum frequencies, respectively. Substitute Equations 4.5 through 4.8 into Equation 4.4 and simplify into linear algebra form:

$$\begin{bmatrix} 1 + \mathcal{S}_{4\pi\hat{f}\xi} & 0 \\ 0 & 1 - \mathcal{S}_{4\pi\hat{f}\xi} \end{bmatrix} \begin{Bmatrix} \hat{a} \\ \hat{b} \end{Bmatrix} = \begin{Bmatrix} (\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi}) \\ 0 \end{Bmatrix} \qquad 4.9$$

The uncoupled solutions are found by inspection:

$$\hat{a} = \frac{\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi}}{1 + \mathcal{S}_{4\pi\hat{f}\xi}} \qquad 4.10$$

$$\hat{b} = 0$$

For this first simple case of zero phase cosine and symmetrical integration limits, we note three results:

- the out-of-phase amplitude is always zero (no surprise since it is orthogonal);
- the estimated in-phase amplitude is biased by errors in the frequency estimate (recall the true amplitude is 1.0); and
- the estimated phase $\hat{\theta} = \tan^{-1}\left(\hat{b}/\hat{a}\right)$ is unbiased and is independent of the integration interval.

This third point is the key. Since $\hat{b}$=0, then $\hat{\theta}$=0, and most importantly, $\hat{\theta} = \theta$ always.

This result is equivalent if a sine instead of a cosine is used for the signal. Note that the solution for this case follows the previous case except for slightly-modified right hand side integrals:

$$\begin{bmatrix} 1 + S_{4\pi\hat{f}\xi} & 0 \\ 0 & 1 - S_{4\pi\hat{f}\xi} \end{bmatrix} \begin{Bmatrix} \hat{a} \\ \hat{b} \end{Bmatrix} = \begin{Bmatrix} 0 \\ S_{2\pi f_\Delta \xi} - S_{2\pi f_\Sigma \xi} \end{Bmatrix} \qquad 4.11$$

The solution is again easily seen by inspection:

$$\hat{a} = 0$$

$$\hat{b} = \frac{S_{2\pi f_\Delta \xi} - S_{2\pi f_\Sigma \xi}}{1 - S_{4\pi\hat{f}\xi}} \qquad 4.12$$

The conclusions are unchanged from the cosine case; the orthogonal amplitude is zero, so the estimated phase will always equal the true phase.

These first two examples were considered separately to illustrate this important phase matching behavior as clearly as possible. The next

logical step is to analyze a monochromatic sinusoid with arbitrary phase which is simply the linear sum of these previous two special cases. Start by defining the true signal with independent in-phase and out-of-phase amplitudes.

$$x(t) \equiv a\cos(2\pi ft) + b\sin(2\pi ft) \qquad 4.13$$

This form corresponds to the model form in Equation 4.2. The matrix equation to solve for the two estimated coefficients is seen to be the same on the left hand side as Equation 4.9 and 4.11, with the scaled sum of the two right hand side vectors from those two equations:

$$\begin{bmatrix} 1+\mathcal{S}_{4\pi\hat{f}\xi} & 0 \\ 0 & 1-\mathcal{S}_{4\pi\hat{f}\xi} \end{bmatrix} \begin{Bmatrix} \hat{a} \\ \hat{b} \end{Bmatrix} = \begin{Bmatrix} a\left(\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi}\right) \\ b\left(\mathcal{S}_{2\pi f_\Delta \xi} - \mathcal{S}_{2\pi f_\Sigma \xi}\right) \end{Bmatrix} \qquad 4.14$$

The two solutions are seen to be functions of the two true amplitudes:

$$\hat{a} = a\left(\frac{\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi}}{1+\mathcal{S}_{4\pi\hat{f}\xi}}\right)$$

$$\hat{b} = b\left(\frac{\mathcal{S}_{2\pi f_\Delta \xi} - \mathcal{S}_{2\pi f_\Sigma \xi}}{1-\mathcal{S}_{4\pi\hat{f}\xi}}\right) \qquad 4.15$$

Next, use this information to examine the behavior of the estimated phase. First, note that it was shown in the two previous special cases that the $\hat{a}$ and $\hat{b}$ coefficient estimates do vary proportionally to the frequency error

and the integration length, so there is a need to closely examine how the estimated phase behaves for this arbitrary phase case. But second, since this signal is a linear sum of the two orthogonal components, there is reason to expect that the estimated phase will still correctly match the true phase. This can be checked by equating the arguments of the estimated and true tangent phase functions from Equation 4.15:

$$\arg(\hat{\theta}) \stackrel{?}{=} \arg(\theta)$$

$$\frac{\hat{b}}{\hat{a}} \stackrel{?}{=} \frac{b}{a} \qquad 4.16$$

$$\left(\frac{\hat{b}}{\hat{a}}\right)\left[\left\{\frac{(S_{2\pi f\Delta\xi} + S_{2\pi f\Sigma\xi})}{(1+S_{4\pi\hat{f}\xi})}\right\}\left\{\frac{(1-S_{4\pi\hat{f}\xi})}{(S_{2\pi f\Delta\xi} - S_{2\pi f\Sigma\xi})}\right\}\right] \stackrel{?}{=} \frac{b}{a}$$

This latter expression shows that for the estimated phase to always be equal to the true phase, then the bracketed term must be 1.0 for all sets of frequencies and integration periods. In other words, while $\hat{a}$ and $\hat{b}$ are known generally to both be incorrect (biased), their ratio must remain invariant. The behavior of this equation and the equations for $\hat{a}/a$ and $\hat{b}/b$ is investigated in more detail in Appendix B. That investigation does show that the bracketed term above does stay approximately equal to 1.0 as required to satisfy Equation 4.16. This in turn establishes the important point that *it is possible to apply an estimated (biased) frequency to an unknown monochromatic signal and recover at least one true (unbiased) parameter - the phase.*

4.3 Geometric Interpretation of HPT as Applied to a Single Sinusoid

The algebraic conclusions from the previous section of fitting a single sinusoid with an estimated frequency to a single sinusoid with an unknown frequency can be reinforced with a geometrical interpretation. Recall the least squares estimator from Equation 4.3 and rewrite it as two sinusoids with arbitrary amplitudes and phases:

$$\begin{aligned} Q &= \int_{-\xi}^{\xi} [\hat{x}(t) - x(t)]^2 \, dt \\ &= \int_{-\xi}^{\xi} [\hat{c}\cos(2\pi\hat{f}t + \hat{\vartheta}) - c\cos(2\pi f t + \vartheta)]^2 \, dt \end{aligned} \qquad 4.17$$

When the true and estimated frequencies are close, then $\hat{c} \approx c$, and the integrand can be equivalently represented as the product of an instantaneous sinusoid with a slowly-varying modulating envelope as outlined in Appendix A:

$$Q \cong 2c \int_{-\xi}^{\xi} [\cos(2\pi f_{\Delta} t + \gamma)\cos(2\pi \bar{f} t + \mu)]^2 \, dt \qquad 4.18$$

where $f_{\Delta} \equiv \left| \hat{f} - f \right| / 2$, $\bar{f} \equiv (\hat{f} + f)/2$, and the inequality comes from the inequality in the amplitudes; the phases in these last two equations are not relevant and are only included for completeness.

The least squares error function Q in Equation 4.18 can be interpreted as the area under a rectified (squared) beating sinusoid. Furthermore, since

it is assumed here that $\hat{c} \approx c$, then the best-fit estimator for $\hat{\vartheta}$ (in Equation 4.17) minimizes that positive Q area. For any fixed time interval, adjusting the estimated phase $\hat{\vartheta}$ changes the interaction between the estimated and true sinusoids, and in the process effectively shifts the modulating envelope of the underlying instantaneous sinusoid in time. So an equivalent geometrical interpretation of the algebraic least squares problem is to shift the estimated phase until the area under the modulated error signal is a minimum over some fixed integration interval $t_c - \xi \leq t \leq t_c + \xi$ where t_c is the center time of the region. It is intuitive to state that the minimum area for this semi-positive definite product integrand function will always be centered about a minimum of the absolute value of the envelope (or equivalently, a node of the zero mean envelope). In fact, this minimum is reduced to zero when $\hat{c} \approx c$ (thereby defining the zero-crossing of the envelope at a time denoted as t_{zc}). Finally, by recognizing from Equation 4.17 that the envelope of the Q function is zero only when $\hat{x}(t_{zc}) = x(t_{zc})$, it is seen that the estimated and true phases must be equal for this to occur since the $2\pi f t_{zc}$ contribution is zero. [Note also that the envelope of Q is a maximum when $\hat{x}(t_{zc}) = -x(t_{zc})$.] This geometrical interpretation is illustrated in Figures 4.1a and 4.1b.

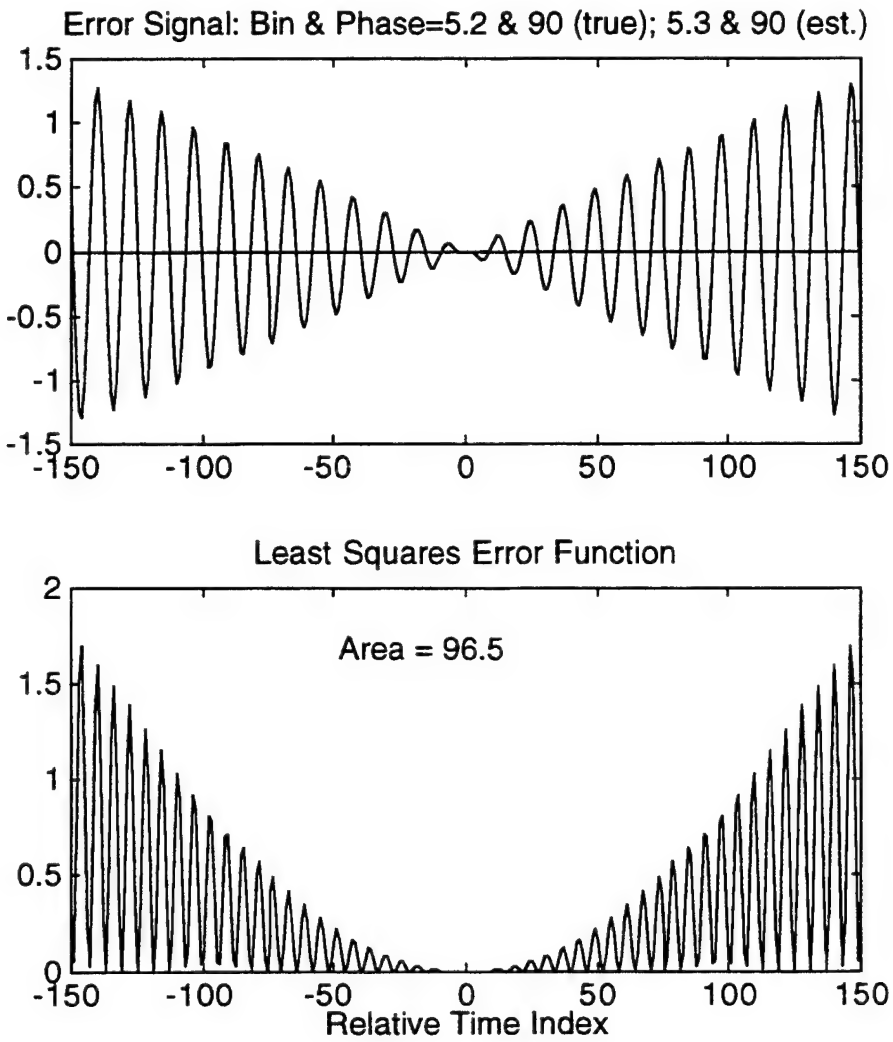

Figure 4.1a. Minimum HPT Error Signal when Estimated Phase Equals the True Phase

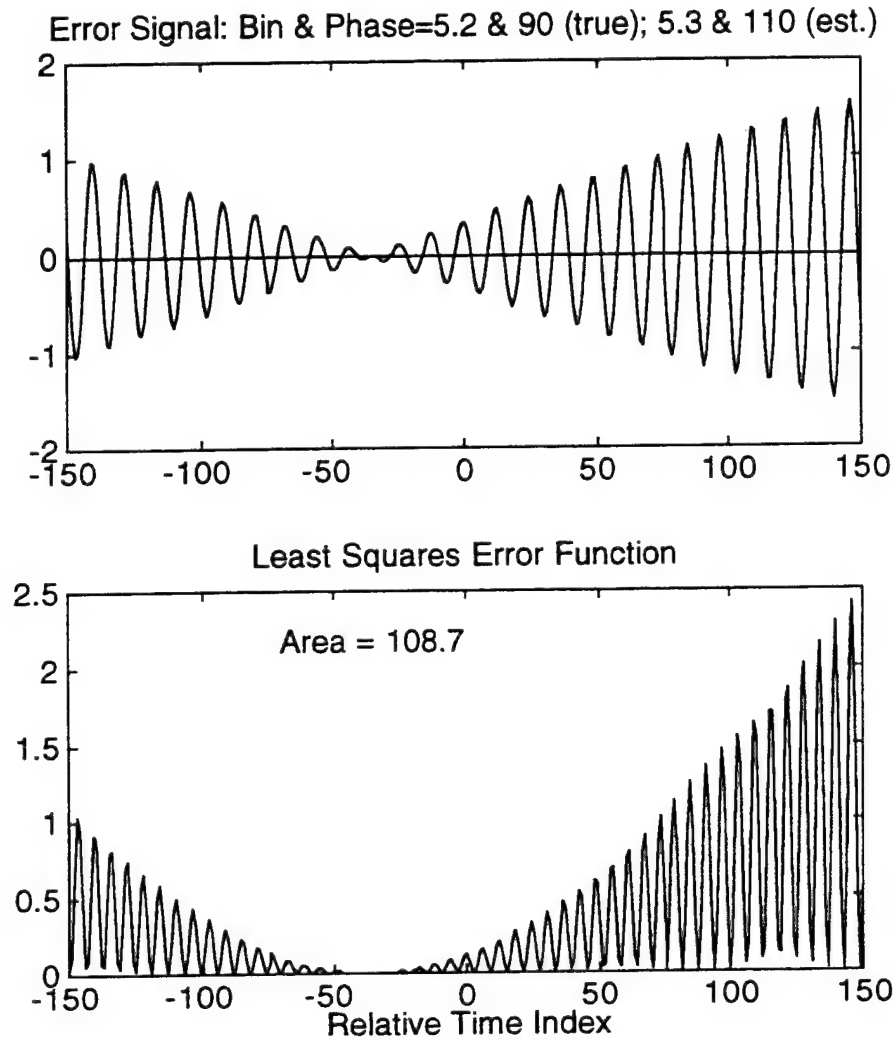

Figure 4.1b. Non-minimum HPT Error Signal when Estimated Phase Does Not Equal the True Phase

In both plots the second subplot is the error function integrand in Equation 4.17. Figure 4.1a illustrates the correct case in which the estimated phase equals the true phase, even for an incorrect frequency. In Figure 4.1.b the estimated phase was purposely shifted by 20 degrees to illustrate how the area of the error function varies from the minimum value in Figure 4.1.a.

A second geometrical way of visualizing this phase matching process is to consider the case where the estimated amplitude and the estimated phase (but not the frequency) are both unbiased. For time equals zero the frequency contribution to the trigonometric arguments will be zero for both sinusoids (see Equations 4.2 and 4.13); as a result the estimated and true signals will be equal and the error will be zero. However, since the frequencies are different, an error will grow with an envelope (that initially increases linearly) as the two unequal frequencies contribute different amounts versus time to each cosine argument. As above, the minimum error area must straddle this time value where the error is zero - namely, where the phases are zero.

4.4 Review: HPT Phase Tracking and Frequency Correction as Applied to a Single Sinusoid Signal

Sections 4.2 and 4.3 showed that a least squares technique will return an unbiased estimate of the true phase for any segment of a single sinusoid deterministic signal as long as the estimated frequency is close to the true frequency. *The key to the Harmonic Phase Tracking technique is that this behavior can be exploited to identify the true frequency.*

The approach begins by identifying one section of the time series. Recall that an estimated (trial) frequency is assumed known, and define a "local" zero-mean time variable centered at reference time t_c^{ref}. Perform a least squares fit (as described in Section 4.2) to estimate the "effective" phase defined from the coefficients as

$$\hat{\Omega} \equiv 2\pi\hat{f}t + \hat{\vartheta} = \tan^{-1}\left(\hat{b}/\hat{a}\right) \cong \Omega \qquad 4.19$$

Since the local time is zero-mean, then this estimated effective phase will equal the true constant phase value ϑ (i.e., using a zero-mean time vector eliminates possible bias due to errors in $\hat{f}$ from the $2\pi\hat{f}t$ component). Next, shift the time series segment forward (say, a small percentage of the segment length and typically keep the same segment length), redefine a new local zero-mean time vector and center time t_c^j, and re-estimate the "effective" phase relative to that new center time. Repeat this process several times in the forward direction and a similar number in the

negative direction. This defines two numerical functions: one with the shift in center times ($\Delta t_c \equiv t_c^j - t_c^{ref}$) and one with phase estimates $\hat{\Omega}(t_c)$.

Next, return to the concept of Instantaneous Frequency discussed in Appendix A, Section A.2. Note that Equation A.12 shows that the frequency can be found from the time derivative of the "effective" phase function defined above:

$$\frac{d\Omega}{dt} = \frac{d}{dt}(2\pi ft + \vartheta) = 2\pi f \qquad 4.20$$

By making the numerical approximation that

$$\frac{d\Omega}{dt} \approx \frac{\Delta\Omega}{\Delta t_c} \qquad 4.21$$

The true frequency is readily estimated by equating these two equations:

$$f \approx \frac{1}{2\pi}\left(\frac{\Delta\Omega}{\Delta t_c}\right) \qquad 4.22$$

Geometrically, when $\Omega(t_c)$ is plotted versus Δt_c, the slope is $2\pi f$ - *not* $2\pi\hat{f}$. *That is the key observation - the* <u>true</u> *frequency is recoverable directly from use of an* <u>*estimated*</u> *frequency.*

The process described in this section and resulting in Equation 4.22 for determining the unknown frequency of an arbitrary single sinusoid can be summarized as:

1. Estimate the true frequency (the technique used for this study is outlined in Section 4.7).

2. Estimate the local "effective" phase using least squares.

3. Shift the segment and estimate a new effective phase.

4. Repeat step 2 as many times as desired, shifting forward and/or backwards with any set of shift values to define a phase function and a time shift function (uniformity is not necessary; symmetry about the initial segment is recommended).

5. The (circular) true frequency is found as the slope of the plot of effective phase function versus the time shift function.

Strictly speaking, if the signal was known to be a single pure sinusoid with no noise, then one time shift and one new phase estimate are sufficient to identify the slope and in turn the true frequency.

The final necessary step in fitting the signal is to estimate the amplitude and phase. These are available from performing one last least squares analysis of the center segment using this "best" frequency.

4.5 Effect of Additive Noise in a Single Sinusoidal Signal.

Noise immediately converts this deterministic problem into a stochastic problem. Noise will definitely add a variance and in some cases a bias to the phase estimates. The usual remedy for eliminating the variance in stochastic estimates is to algebraically average across independent "raw"

estimates (a good example is ensemble averaging with FFT-based spectral analysis).

The phase tracking technique described and summarized in the last section works for stochastic as well as a deterministic signals. Consider the usual definition of noise as uncorrelated and white. The presence of such noise will certainly add variance to the effective phase estimate function. But fitting a linear slope automatically acts to smooth the estimated phases and thereby minimize the noise-induced bias. Secondly, since the time shifts are recommended to be small percentages of the segment length, then adjacent phase estimates use redundant data and are therefore correlated, meaning that any noise effects will likewise be shared between all the estimates; this would introduce a [slowly-varying] bias to the phase function, but since it has a slowly-varying magnitude it would not greatly affect the slope of the phase versus time plot. Hence, the process described for the single deterministic sinusoidal signal is equally appropriate when stochastic noise is added. In other words, the noise would add both bias and variance to the $\Omega(t_c)$ function, but both are minimized when the first order slope is fitted. As mentioned, the least squares error function is an integrator typically over several cycles of the signal, and integration is a smoothing operation that will further act to minimize the variance in each of the shifted phase estimates due to noise. This shows that the estimated slope fitted to many shifted phases should still be a reliable estimator even for the case with noise.

The previous section did state that, for a pure sinusoidal signal, that only one time shift was required. That would produce a very unreliable phase estimate if noise was present. The simplest way then to accomodate noise is to use the same deterministic signal process but increase the number and range of the time shifts to provide more information for the slope estimation (which in turn requires a longer segment of the data vector).

Thus, these last three sections have outlined a robust technique for estimating the frequency, amplitude, and phase of a *single* sinusoid with or without additive noise.

The next section extends this technique to a *multiharmonic* signal defined as a finite summation of sinusoids with additive noise.

4.6 HPT Algebraic Development as Applied to a Multiharmonic Signal

The technique presented in the previous sections for identifying the true frequency of a single sinusoid, with and without noise, is readily extended to the case of a multiharmonic signal with and without noise.

The objective is to estimate the frequencies and corresponding amplitudes and phases of a multiharmonic signal s(t) with uncorrelated noise n(t) defined as:

$$\begin{aligned} x(t) &\equiv s(t)+n(t) \\ &\equiv \sum_{j=1}^{r} c_j \cos(2\pi f_j t+\theta_j)+n(t) \\ &\equiv \sum_{j=1}^{r} \left[a_j \cos(2\pi f_j t)+b_j \sin(2\pi f_j t)\right]+n(t) \end{aligned} \qquad 4.23$$

In general, this true number of sinusoids (r) is unknown, and accurate identification of r is in fact often an important part of the analysis. In some instances, it is apparent upon inspection of the results that only a certain number of components have significant amplitudes, and that number can be taken as an estimate of r. However, for many geophysical signals, there is no clear demarcation in amplitudes, or, the objective of the analysis is in fact to quantify small amplitude nonlinear components (such as coupled Stokes harmonics); in these cases the estimation of r is more difficult, and it can in fact vary among analyses. Jammalamadaka and Sama (1993) present an approach using circular regression that

presents a statistical approach for finding r. This topic is revisited in later Chapters.

Define the estimated signal (i.e., the model) as:

$$\begin{aligned}\hat{s}(t) &\equiv \sum_{j=1}^{\hat{r}} \hat{c}_j \cos(2\pi \hat{f}_j t + \hat{\theta}_j) \\ &\equiv \sum_{j=1}^{\hat{r}} \left[\hat{a}_j \cos(2\pi \hat{f}_j t) + \hat{b}_j \sin(2\pi \hat{f}_j t)\right]\end{aligned} \qquad 4.24$$

Note that all of the parameters in Equation 4.24 are estimates, including the number of sinusoids (r) in the true signal. As with the presentation used in the previous sections, assume first for illustrative purposes that a vector of approximately-correct estimated frequencies ($\hat{\mathbf{f}} \approx \mathbf{f} = \mathbf{f} + \mathbf{f}_\Delta$) is available where $\hat{\mathbf{f}} = \left[\hat{f}_1 \quad \hat{f}_2 \quad \cdots \quad \hat{f}_{\hat{r}}\right]^T$. (Again, a procedure for generating this initial frequency estimate vector is the topic of Section 4.7.)

Begin with the usual least squares approach to find the $\hat{\mathbf{a}} = \left[\hat{a}_1 \quad \cdots \quad \hat{a}_{\hat{r}}\right]^T$ and $\hat{\mathbf{b}} = \left[\hat{b}_1 \quad \cdots \quad \hat{b}_{\hat{r}}\right]^T$ vectors from Equation 4.23. Define an error function as the integral of the squared error over some time integration interval:

$$\begin{aligned}Q &= \int_{-\xi}^{\xi} [\hat{s}(t) - x(t)]^2 \, dt \\ &= \int_{-\xi}^{\xi} \left\{ \sum_{j=1}^{\hat{r}} \left[\hat{a}_j \cos(2\pi \hat{f}_j t) + \hat{b}_j \sin(2\pi \hat{f}_j t)\right] - x(t) \right\}^2 dt\end{aligned} \qquad 4.25$$

As before, take partial derivatives with respect to the individual components in both $\hat{a}$ and $\hat{b}$; for example, a sample partial derivative is:

$$\frac{\partial Q}{\partial \hat{a}_i} = 2\int_{-\xi}^{\xi}\left\{\sum_{j=1}^{\hat{r}}\left[\hat{a}_j\cos(2\pi\hat{f}_j t)+\hat{b}_j\sin(2\pi\hat{f}_j t)\right]-x(t)\right\}\cos(2\pi\hat{f}_i t)\,dt \qquad 4.26$$

Set this equal to zero and rearrange:

$$\int_{-\xi}^{\xi}\left\{\sum_{j=1}^{\hat{r}}\left[\hat{a}_j\cos(2\pi\hat{f}_j t)+\hat{b}_j\sin(2\pi\hat{f}_j t)\right]\right\}\cos(2\pi\hat{f}_i t)\,dt = \int_{-\xi}^{\xi}x(t)\cos(2\pi\hat{f}_i t)dt \qquad 4.27$$

Interchange the linear summation and integral operators:

$$\sum_{j=1}^{\hat{r}}\left[\hat{a}_j\int_{-\xi}^{\xi}\cos(2\pi\hat{f}_j t)\cos(2\pi\hat{f}_i t)dt+\hat{b}_j\int_{-\xi}^{\xi}\sin(2\pi\hat{f}_j t)\cos(2\pi\hat{f}_i t)dt\right] = \int_{-\xi}^{\xi}x(t)\cos(2\pi\hat{f}_i t)dt \qquad 4.28$$

Since the frequencies on the left hand side of Equation 4.28 and the complimentary equations for the other $\hat{\mathbf{a}}$ and $\hat{\mathbf{b}}$ coefficients are known, those integrals can be analytically evaluated just as with the single sinusoid case previously presented in Sections 4.2 to 4.5. Recall that in the single sinusoid case, the two (equal) off-diagonal terms in the basis matrix were identically zero and the in- and out-of-phase coefficients were uncoupled. However, Equation 4.28 has multiple arbitrary frequencies over arbitrary (but symmetric) integration intervals, so it is expected that several off-diagonal terms in the basis matrix will be non-zero.

The problem formulation for this multiharmonic case can be illustrated by assuming two sinusoids and using the analytical integrals outlined in Section 4.4. After simplification and rearrangement, Equation 4.28 becomes:

$$\begin{bmatrix} 1+\mathcal{S}_{4\pi\hat{f}_1\xi} & \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}+\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} & 0 & 0 \\ \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}+\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} & 1+\mathcal{S}_{4\pi\hat{f}_2\xi} & 0 & 0 \\ 0 & 0 & 1-\mathcal{S}_{4\pi\hat{f}_1\xi} & \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}-\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} \\ 0 & 0 & \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}-\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} & 1-\mathcal{S}_{4\pi\hat{f}_2\xi} \end{bmatrix} \begin{Bmatrix} \hat{a}_1 \\ \hat{a}_2 \\ \hat{b}_1 \\ \hat{b}_2 \end{Bmatrix} = \begin{Bmatrix} \int_{-\xi}^{\xi} x(t)\cos(2\pi\hat{f}_1 t)dt \\ \int_{-\xi}^{\xi} x(t)\cos(2\pi\hat{f}_2 t)dt \\ \int_{-\xi}^{\xi} x(t)\sin(2\pi\hat{f}_1 t)dt \\ \int_{-\xi}^{\xi} x(t)\sin(2\pi\hat{f}_2 t)dt \end{Bmatrix} \qquad 4.29$$

where $f_{\Delta_{12}} \equiv \left|f_2 - f_1\right|$ and $f_{\Sigma_{12}} \equiv \left|f_2 + f_1\right|$.

Inspection of the coefficient vector shows that the ordering of Equation 4.29 is such that the estimated cosine basis vectors are grouped in the first two columns of the basis matrix, while the second two represent the

estimated sine basis vectors. Thus, the basis matrix can be thought of as divided into a two-by-two partition substructure. The two cosine and two sine integrals in the upper left and lower right partitions of the basis matrix are non-zero since the integrand is an even function in both cases and the limits are symmetrical. The upper right and lower left partitions are always zero since the integrands are odd [cosine-sine] functions over symmetrical limits. This means that Equation 4.29 can be reduced to two partition equations, one for the in-phase and one for the out-of-phase coefficients:

$$\begin{bmatrix} 1+\mathcal{S}_{4\pi\hat{f}_1\xi} & \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}+\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} \\ \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}+\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} & 1+\mathcal{S}_{4\pi\hat{f}_2\xi} \end{bmatrix} \begin{Bmatrix} \hat{a}_1 \\ \hat{a}_2 \end{Bmatrix} = \begin{Bmatrix} \int_{-\xi}^{\xi} x(t)\cos(2\pi\hat{f}_1 t)dt \\ \int_{-\xi}^{\xi} x(t)\cos(2\pi\hat{f}_2 t)dt \end{Bmatrix} \qquad 4.30a$$

and

$$\begin{bmatrix} 1-\mathcal{S}_{4\pi\hat{f}_1\xi} & \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}-\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} \\ \mathcal{S}_{2\pi f_{\Delta_{12}}\xi}-\mathcal{S}_{2\pi f_{\Sigma_{12}}\xi} & 1-\mathcal{S}_{4\pi\hat{f}_2\xi} \end{bmatrix} \begin{Bmatrix} \hat{b}_1 \\ \hat{b}_2 \end{Bmatrix} = \begin{Bmatrix} \int_{-\xi}^{\xi} x(t)\sin(2\pi\hat{f}_1 t)dt \\ \int_{-\xi}^{\xi} x(t)\sin(2\pi\hat{f}_2 t)dt \end{Bmatrix} \qquad 4.30b$$

This partitioning saves considerable computer time for problems with a large number of sinusoids in the signal. For discrete data, the integrals on the right hand side of Equations 4.30 are approximated by summations; for example:

$$\int_{-\xi}^{\xi} x(t)\cos(2\pi\hat{f}_1 t)dt = \Delta t\left[\frac{z_1+z_{L+1}}{2}+\sum_{j=2}^{L} z_j\right] \qquad 4.31$$

where $z_j \equiv x_j \cos(2\pi \hat{f}_1 t_j)$ and L+1 is the length of the discrete x_j data vector such that $\xi = L/2$.

The final step is to solve Equations 4.30. Note that the basis matrix is only an estimate of the true basis matrix since the true frequencies are not known (in particular, the off-diagonal terms can have significant errors), and in fact change during the analysis. For this reason it is *incorrect* to solve these equations using conventional least squares, and instead total least squares is used.

For applications where noise is present, N equations are needed to exactly fit N data points regardless of the size of r. In these general cases Equations 4.30 are overdetermined and the equations are not consistent. Since there is no exact solution a minimum norm measure must be used instead. Consider representing Equations 4.29 in the usual $\mathbf{Ax}=\mathbf{d}$ form. In conventional least squares, the solution is found from the normal equations given by

$$\mathbf{A}^T\mathbf{A}\mathbf{x} = \mathbf{A}^T\mathbf{d} \qquad 4.32$$

or equivalently

$$\mathbf{A}^T\mathbf{e} = 0 \qquad 4.33$$

where $\mathbf{e} \equiv \mathbf{Ax} - \mathbf{d}$ is the error vector. Equation 4.33 shows that $\mathbf{e}$ must be orthogonal to the range of $\mathbf{A}^T$. Thus, in conventional least squares the "best" estimated solution is defined as the vector *in the range of* A^T such

that the vector sum of it and **e** equal the original **d** vector. This least squares vector solution is quantitatively defined by:

$$\mathbf{x}_{ls} \equiv \left(\mathbf{A}^T\mathbf{A}\right)^{-1}\mathbf{A}^T\mathbf{d} \qquad 4.34$$

The important point here is that it is implicit in conventional least squares that the basis matrix is exact since the solution is forced into its range. However, the basis matrix in Equation 4.30 is constructed from estimated frequencies and is *not* exact, so a more appropriate solution technique to apply is *total* least squares which is described briefly next.

Total least squares essentially geometrically recasts the problem before taking advantage of the properties of the least squares estimator described above (Van Huffel and Vandewalle, 1991). Simply rearrange the least squares formulation to define a new "augmented basis matrix" and a correspondingly modified solution vector:

$$\left[\mathbf{A} \quad \vdots \quad \mathbf{d}\right]\begin{Bmatrix}\mathbf{x}\\ -1\end{Bmatrix} = \mathbf{0} \qquad 4.35$$

This is algebraically equivalent to saying $\mathbf{Ax}-\mathbf{d}=\mathbf{0}$, but geometrically it is quite different. The fundamental difference is that the basis matrix has been altered to incorporate the data column vector **d** (resulting in a coefficient vector that is one longer than before). This means that the range of the transpose of this new basis matrix is affected by information in the **d** vector as well as the **A** matrix. The total least squares solution to Equation 4.35 is equal to the null vector of the transpose of this new basis

matrix. As previously stated, whenever noise is present then this new augmented matrix is full rank.

In practice, the recommended technique for solving Equation 4.35 starts by performing a singular value decomposition of $[\mathbf{A} \vdots \mathbf{d}]$. By the minimum norm properties of the singular value decomposition, the vectors in the left and right singular matrices (**U** and **V**, respectively) are ordered in terms of decreasing contribution to the $[\mathbf{A} \vdots \mathbf{d}]$ matrix (as quantified by the singular values). Therefore, the last right singular value contributes the least and is accordingly it is the best estimate of the null vector required to solve Equation 4.35. Since this is an overdetermined system of equations when noise is present, the right singular matrix will generally be full rank and thus does yield a viable last right singular vector for estimating $[\mathbf{x}^T \ -1]^T$. But because the singular vectors are orthonormalized, a final scaling is required to set the last element of this null vector to -1 (note that this does not affect the solution to Equation 4.35 since it is homogeneous); then, the remaining scaled elements are equal to the total least squares estimate for **x**.

Summarizing: given an estimated basis matrix with dimensions $\mathbf{A}_{MxN}$ with M>N, and a data vector $\mathbf{d}_{Nx1}$. Construct a new matrix $[\mathbf{A} \vdots \mathbf{d}]_{Mx(N+1)}$. Since it is assumed that no exact solution is available (i.e, **d** is not in the range of $\mathbf{A}^T$), then the new matrix is full rank. The "best" rank N matrix approximation to this new rank N+1 matrix (i.e., a consistent matrix

representing the range of $\mathbf{A}^T$ and $\mathbf{d}$) is found from the summation of the first N singular values and vectors from a singular value decomposition. If this new consistent matrix is denoted as $[\mathbf{A} \vdots \mathbf{d}]$, then the best total least squares solution can be shown to be:

$$\hat{\mathbf{x}}_{tls} = \hat{\mathbf{A}}^{-1}\hat{\mathbf{d}} \qquad 4.36$$

However, in practice a different definition is used. The remaining right singular vector $\mathbf{v}_{N+1}$ not used in the rank-N approximation $[\mathbf{A} \vdots \mathbf{d}]$ is orthogonal to this new best matrix and therefore represents an approximation to the null row space. So instead, directly scale this vector to arrive at the best total least squares estimate for the coefficient vector:

$$\hat{\mathbf{x}}_{tls} \equiv -\frac{1}{v_{N+1,N+1}}\left[v_{1,N+1} \quad , \quad \cdots\cdots \quad , \quad v_{N,N+1}\right]^T \qquad 4.37$$

Figure 4.2a and 4.2b illustrate the geometrical difference between the conventional and total least squares problems - primarily, that the range of the basis matrix is different.

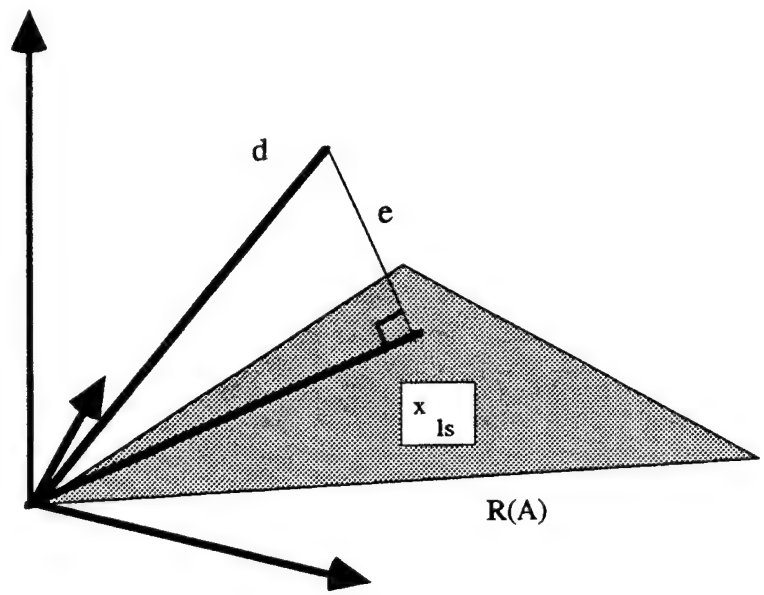

Figure 4.2a Least Squares (LS) Geometry

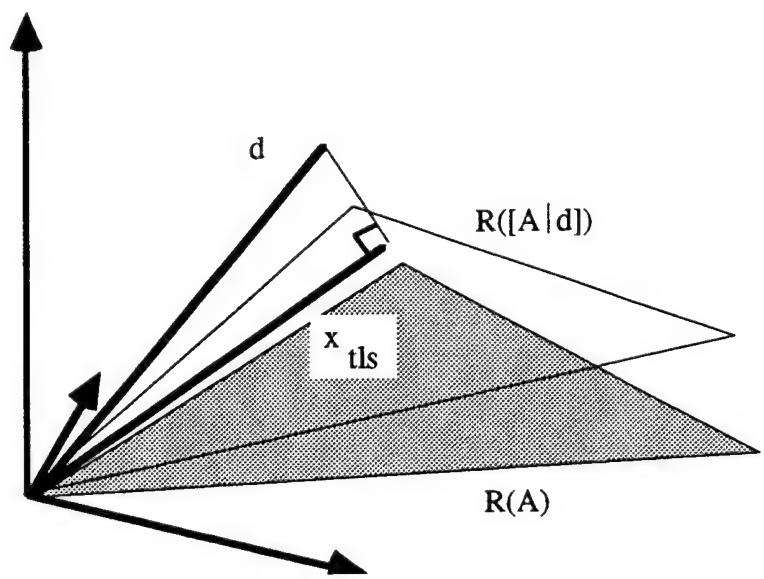

Figure 4.2b Total Least Squares (TLS) Geometry

It has been shown that the total least squares estimator is significantly more robust for almost-collinear multicollinearities (i.e., ill-conditioned problems) compared to conventional least squares (otherwise, the solutions are similar). As discussed in later Chapters, this is a valuable characteristic for signal processing applications such as the new Harmonic Phase Tracking technique developed here.

The preceding presentation in this section has described the solution methodology for finding the best estimate of the coefficient vector, relative to one estimate of the frequency vector $\hat{\mathbf{f}}=\left[\hat{f}_1\ \hat{f}_2\ \cdots\ \hat{f}_{\hat{r}}\right]$, and for one particular segment from the data realization:

- Use this frequency vector to define the *analytical* basis matrix $\hat{\mathbf{A}}$ and *numerical* data vector $\hat{\mathbf{d}}$ using Equations 4.29 and 4.30;
- perform a singular value decomposition on the augmented, estimated basis matrix $\left[\mathbf{A}\ \vdots\ \mathbf{d}\right]$;
- estimate the total least squares coefficient vector $\hat{\mathbf{x}}_{tls}=\left[\hat{\mathbf{a}}\ \vdots\ \hat{\mathbf{b}}\right]^T=\left[\hat{a}_1\ \hat{a}_2\ \cdots\ \hat{a}_{\hat{r}}\ \vdots\ \hat{b}_1\ \hat{b}_2\ \cdots\ \hat{b}_{\hat{r}}\right]^T$ using Equation 4.37; note that the estimated rank may not be equal to the true rank of the signal.
- convert the coefficient vector to a phase vector referenced to the center time of this segment:

$$\hat{\Omega}(t_c)=\tan^{-1}\left(\hat{\mathbf{b}}/\hat{\mathbf{a}}\right) \qquad 4.38$$

Equation 4.38 is simply a vector version of Equation 4.19 (i.e., the argument is intended in an inner product sense). And, again, the trigonometric function must account for the proper quadrant of the phase.

The next step is to shift the fixed-length segment forward and backward, then for each shift to recalculate the right hand side $\hat{\mathbf{d}}$ vector and estimate corresponding phase vectors using Equation 4.38.

At this point the solution methodology diverges from a simple scalar-to-vector extension of the single sinusoid case. It would seem that the true [multiple] frequencies could be estimated following scalar Equation 4.21:

$$f_k \overset{?}{=} \frac{1}{2\pi}\left(\frac{\Delta\Omega_k}{\Delta t_c}\right) \qquad k = 1, 2, \cdots, \hat{r} \qquad 4.39$$

Unfortunately, estimates using Equation 4.39 are not always accurate. The source of the problem lies in the fact that, in the multiharmonic case, the off-diagonal terms in the basis matrix are not zero, so that the coefficients are coupled and biased. Those terms were zero for the single sinusoid case (where an error in the estimated frequency does admittedly introduce a [very small] bias in the single phase analytical function defined in Equation 4.16). In the multiharmonic case, errors are introduced into the coefficient vector estimate, and hence into the phase vector estimate, whenever biased frequencies are used in the basis matrix (or noise is

present in the signal). Those errors further illustrate the value of total rather than conventional least squares as the preferred solution method.

So, direct application of Equation 4.39 is not used because it was found to be unreliable. All of this indicates that an iterative technique is appropriate for the multiharmonic case. During the nth iteration, the frequency vector from the previous iteration, denoted $\hat{\mathbf{f}}^{(n-1)}$, is increased or decreased proportional to a modified factor based on Equation 4.39 to become the new estimate $\hat{\mathbf{f}}^{(n)}$. Consider each ratio of the updated versus the initial estimated frequencies $\hat{\mathbf{f}}_k^{(n)} / \hat{\mathbf{f}}_k^{(n-1)}$ at the end of this n^{th} iteration. While these ratios are not always *absolutely* correct, it has been found that the *relative* magnitude of this ratios are typically correct (i.e., large versus small increase, or large versus small decrease).

The frequency adjustment technique at each iteration, as implemented in the computer code, starts with the apparent adjustment to the initial estimate of frequency indicated by the $\hat{\mathbf{f}}_k^{(n)} / \hat{\mathbf{f}}_k^{(n-1)}$ ratio, but modifies it based on several other factors. Therefore, this iterative stage of the solution process cannot be described as rigorous, and because of that it can be implemented in any number of ways. It has been found that the estimated frequency vector can exhibit quite dynamic behavior during this iterative process. Example situations illustrating that dynamic behavior include:

- *frequency removal:* in some cases the amplitudes at a given frequency will asymptotically go to zero. This occurs in situations where an initial (or inserted) frequency is incorrect. When such small amplitudes are detected that frequency is removed from the frequency vector as noncontributory, and, to reduce the size of the basis matrix.

- *frequency removal:* in some cases where there are two closely-spaced true frequencies (say, just below the resolution limit of this technique), the initial estimated frequency vector may contain two frequencies spaced slightly wider. The iterative technique will result in both frequencies [correctly] converging towards the true values, but if they get too close the analytical coupling in the basis matrix increases the condition number too much, at which time the two estimated frequencies are purposely replaced with one equivalent frequency (the time-frequency ambiguity problem).

- *frequency addition:* the estimate for the frequency vector leads to estimation of the coefficient vector, and from that the time domain error between the data and the present fit can be calculated at each iteration. If the spectrum of that error shows any individual ordinates above a selected threshold, then it is necessary to insert frequencies into the frequency vector in that region.

These few examples demonstrate that the frequency vector is quite adaptive during the iterations, which is actually a powerful characteristic of this Harmonic Phase Tracking technique. Appendix C contains a thorough review of the many algorithms, checks, and thresholds chosen for this implementation.

The final step is to identify measures to indicate convergence of this iterative/asymptotic process. Two measures are used. The first is absolute - if the ratio of the root mean square value of the error divided by the root mean square value of the data vector over the finite segment of interest is less than a threshold - say 0.5 percent - and the estimated amplitudes and phases are not significantly improving with the iterations, then convergence is conditionally assumed.

The second measure is relative convergence, which is the more difficult and more universal of the two. If the noise is not negligible, then the absolute criterion will never be reached with a small number of modeled components. In this more typical case, the iterations asymptotically converge to a "best" case where no further improvements are evident. As implemented, convergence is assumed if the maximum variations in the last several amplitude, phase, and frequency vectors all are less than a selected threshold (subjectively chosen based on many numerical studies).

Strictly speaking, since the iterative process asymptotically converges, the technique never fully achieves "the exact answer" even in deterministic validations versus known signals. However, it should not be inferred from this statement that the technique is either slow to converge or only capable of returning approximations; in deterministic (noise-free) validation studies, accuracies of two and often three significant digits for amplitude, phase, and frequency using 10 to 30 iterations are typical. Asymptotic convergence is, of course, also inherent in many other numerical models such as finite and boundary elements where it is necessary and acceptable to define convergence thresholds that are trade-offs between accuracy and computer time. Appendix C also contains full discussions of these issues, and a full summary of the technique as implemented. Numerical performance is addressed there and in the next three Chapters.

4.7 Identification of the HPT Initial Frequency Vector for Multiharmonic Signals

In all of the preceding discussions in this Chapter, it has been assumed that a vector of estimated frequencies was available to begin the iterative process. This section presents a technique for identifying this initial vector of estimated signal frequencies defined as

$$\hat{\mathbf{f}}^{(0)} = \left[\hat{f}_1^{(0)} \quad \hat{f}_2^{(0)} \quad \cdots \quad \hat{f}_{\hat{r}}^{(0)}\right]^T .$$

The objective is to estimate $\hat{\mathbf{f}}^{(0)}$ with the following two attributes: (1) reasonable if not correct rank, and (2) reasonably correct frequency values. Actually, neither of these is strictly necessary since the size of the frequency vector ($\hat{\mathbf{r}}$) is dynamically adjusted during the iterative process; but those dynamics (and hence the number of iterations) can be reduced with a reasonably accurate initially estimated frequency vector.

Consider for illustrative purposes a deterministic signal x(t) defined as a single sinusoid, with: unit amplitude, a frequency at Fourier Series integer bin k, unit time steps, and zero phase (i.e., a cosine). A length-N FFT would result in the discrete complex transform vector:

$$\mathbf{X}(f_j|t_0) = \left(\frac{N}{2}\right)\delta\left(f_j - f_k\right) + i\mathbf{0} \qquad j = 0,1,\cdots N/2 \qquad 4.40$$

where t_0 is the reference start time for the segment used for this FFT, δ is the usual delta function operator, and $i = \sqrt{-1}$ is the imaginary operator . In words, the imaginary transform is identically zero, while the real transform is zero except for a finite delta function at the kth Fourier bin.

Now shift the time series forward one time step. This introduces a time domain phase equal to $2\pi f_k(1)$. Thus, the original cosine signal now must be represented with a cosine and a sine wave such that the square of their amplitudes is one (the original amplitude). The same length-N FFT assigned to this first shifted segment would result in:

$$\mathbf{X}(f_j|t_1)=\left(\frac{N}{2}\right)\left\{\mu+i\sqrt{(1-\mu^2)}\right\}\delta\left(f_j-f_k\right) \qquad j=0,1,\cdots N/2 \qquad 4.41$$

where $\mu<1$ represents the fractional amplitude proportional to the real (in-phase, cosine) component (here, $\mu=\cos(2\pi f(1))$).

Imagine continuing this process for M time steps. First, identify the period for a sinusoid at the kth Fourier bin as the reciprocal of the frequency, or $T_k={}^{N}/_{k}$. Thus, for a time shift of one quarter cycle equal to ${}^{N}/_{4k}$, the shifted signal will be a sine wave with negative unit amplitude and a purely imaginary transform

$$\mathbf{X}(f_j|t_{N/4k})=\mathbf{0}+i\left(\frac{N}{2}\right)\delta\left(f_j-f_k\right) \qquad j=0,1,\cdots N/2 \qquad 4.42$$

The real transform vector is identically zero. After ${}^{N}/_{2k}$ time shifts the original unit amplitude cosine appears as a *negative* unit amplitude cosine with transform:

$$\mathbf{X}(f_j|t_{N/2k})=-\mathbf{X}(f_j|t_0)=-\left(\frac{N}{2}\right)\delta\left(f_j-f_k\right)+i\mathbf{0} \qquad j=0,1,\cdots N/2 \qquad 4.43$$

For a time shift of ${}^{N}/_{k}$ the time series has rotated one full cycle and the transform repeats:

$$\mathbf{X}(f_j|t_{N/k})=\mathbf{X}(f_j|t_0) \qquad j=0,1,\cdots N/2 \qquad 4.44$$

This first cyclical transform pattern repeats every ${}^{N}/_{k}$ time shifts ad infinitum.

Continuing with this first single sinusoid example, now define a matrix assembled using columns taken from the real part ($\Re$) of the transforms in chronological order:

$$\mathbf{R}_{(N/2)xM} \equiv \Re\left\{\left[\mathbf{X}(f_j|t_0)\ \mathbf{X}(f_j|t_1) \cdots \mathbf{X}(f_j|t_M)\right]\right\} \qquad 4.45$$

where N is the length of the transform ($^N/_2$ is the Nyquist bin) and M-1 is the number of transforms (i.e., number of time shifts, which can be forward and/or backward). The kth row of **R** represents the signal passed through a narrow (Fourier) filter centered at bin k. For this preceding example case, **R** would equal:

$$\mathbf{R}_{(N/2)xM} \equiv \left(N/2\right)\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 1 & \mu & \cdots & \zeta \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \equiv \left(N/2\right)\begin{bmatrix} 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ \cos(2\pi f_k t') \\ 0 & 0 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \end{bmatrix} \qquad 4.46$$

where t'=[0 1 ..... M-1] and $\zeta=\cos(2\pi f_k(M-1)/(N/2)$ for all forward shifts.

Since **R** is a chronological ordering of transforms, it is seen to be similar to a two-dimensional spectrum $\mathbf{S}(f|t)$ (oftentimes denoted as $\mathbf{S}(f,t)$ in nonstationary studies) of the signal. However, direct use of the transform for **R** retains signs which are rectified and lost in the $\mathbf{S}(f,t)$ representation. In fact, since the Fourier Transform is a linear operator, linear system theory requires that for deterministic signals each row of **R** will exhibit the same frequency as the associated time domain sinusoidal

signal. Thus, **R** is a linear operator whereas S(f,t) is not. This is an important distinction that is discussed further in this section.

Note that for this case of a single sinusoid with an integer Fourier frequency the kth transform row is a scaled version of the time domain signal; that is, they have identical frequencies and phases. If the deterministic time domain sinusoid had been defined with an arbitrary phase, then the phase of the transform would show that same arbitrary phase because **R** is a linear operator.

Thus, for this simplest case, identification of the frequency, amplitude, and phase of the frequency domain transform sinusoid allows for recovery of the parameters of the underlying time domain sinusoid.

The next simplest deterministic time domain signal to consider is a multiharmonic signal, but with frequencies constrained to equal Fourier Series integer frequencies. Since all of the signals are orthogonal, each row of the resulting **R** matrix is unaffected by the other transforms, and again, the component sinusoids can be recovered identically from each row. Therefore, even for this special multiharmonic case, recovery of the time domain sinusoids is straightforward after identification of the component amplitudes and phases (recall that the frequencies are constrained to the Fourier integer harmonics so are already known) of the

frequency domain sinusoids as defined by rows in **R**. Note that each row of **R** will exhibit a different frequency appropriate for each bin.

This scheme of identifying the components in the time domain signal from analysis of the **R** matrix is less straightforward for nonorthogonal multiharmonic signals. For this case, consider a single sinusoid deterministic signal with an arbitrary frequency *not* equal to one of the Fourier harmonic frequencies - i.e., a frequency corresponding to a "fractional" (non integer) bin number. As was shown previously in Figure 3.1, a finite Fourier transform of this signal leaks energy to all of the bins. However, since this leakage is in phase with the sinusoidal signal, *all of the rows of **R** will still exhibit the same single period as the time domain signal.* If the fractional frequency is closest to the kth Fourier bin, then that row will show the largest amplitude. Unfortunately, the phase of the transform at this k^{th} row is no longer necessarily equal the phase of the time domain sinusoid. This is easily illustrated by considering a signal defined as a single zero-phase cosine (relative to a FFT time axis t'=[0 1 ..... M-1]) with a frequency at the $(k+{}^1/_2)^{th}$ bin. This means that a FFT would see $k+{}^1/_2$ cycles of this signal. Since the FFT interprets the original signal as one period of an infinitely-repeating periodic signal, it falsely interprets the original cosine signal (i.e., even function) to instead be an *odd* function (note that in this example each of these periodic segments ends at a ${}^1/_2$ cycle, or trough, then has a discontinuous jump to a crest to start the subsequent segment; thus x(-t)=-

x(t) at the origin). Because this is an odd function, the FFT will return identically zero values for the real transform - even though the original signal was truly an in-phase cosine. Thus, the transform is completely out-of-phase with the time domain signal. Summarizing, the phase shift of the transform relative to the time domain phase is zero for sinusoids with frequencies equal to Fourier (integer) bins, it is 90 degrees when the frequency is exactly between the bins (at the $^1/_2$ fraction), and it was found to be proportional for other arbitrary fractional values using the expression:

$$\Delta\phi_{f\to t} = -\left(\frac{\pi}{2}\right)\left(1+\mathrm{sign}(\varepsilon)\right)+\varepsilon\pi \qquad -\tfrac{1}{2}\le\varepsilon\le\tfrac{1}{2} \qquad 4.47$$

where: $\Delta\phi_{f\to t}$(rad) is added to the frequency domain phase to recover the time domain phase, and ε is the fractional bin value ($-^1/_2<\varepsilon<^1/_2$). Thus, even with these differences, the phase of the time domain signal is recoverable from the frequency domain transform sinusoid for this non-integer period single sinusoid case.

Converting amplitude from the frequency domain (transform) back to the time domain is not as easy. The ideal relationship (using a Fourier Transform and a continuous signal) between the frequency and time domain amplitudes for the rectangular window would be expected to be:

$$\left(\frac{c_f}{c_t}\right)_{\text{no aliasing}} = \left(\frac{N}{2}\right)\left|\frac{\sin(\pi\varepsilon)}{\pi\varepsilon}\right| \qquad 4.48$$

where the subscripts refer to the frequency and time domains, respectively. The complication is due to aliasing caused by discretizing the continuous signal, which biases the transform magnitudes at all bins, particularly for sinusoids at low bin numbers. The chosen solution was to numerically investigate how the two amplitudes vary with FFT length, bin number, and fractional frequency, rather than analytically developing an expression in the form of a summation that folds ad infinitum around multiples of the Nyquist bin number. This numerical relationship uses Equation 4.48 as the basis, then adds two modifying functions to account for the bin number and fractional frequency:

$$\left(\frac{c_f}{c_t}\right)_{\text{aliasing}} \equiv \left(\frac{c_f}{c_t}\right)_{\text{no aliasing}} \gamma_{\text{bin}} \gamma_{\varepsilon} \qquad 4.49a$$

where

$$\gamma_{\text{bin}} = 2 - \frac{1}{\tanh\left(\frac{\varepsilon}{2k}\right)}$$

$$\gamma_{\varepsilon} = 1 - \frac{\text{sign}(\varepsilon)\left(\frac{\varepsilon}{2k}\right)^6}{k+3} \qquad 4.49b$$

where k is the closest integer bin number and ε is the fractional bin value. Equations 4.49b are non unique, reasonably-accurate functional fits determined during the numerical study.

Equations 4.47 and 4.49 demonstrate that the amplitude and phase for a single time domain sinusoidal signal can be identified from analysis of the **R** matrix even for arbitrary (fractional) frequencies.

The next task is to extend this scheme to identify deterministic time domain signals composed of a summation of sinusoids with fractional (arbitrary) frequencies. Since the **R** matrix is inherently a linear mapping (the defining conditions for linear systems hold even when the aliasing effects are included) of the time domain sinusoids to the frequency domain, this is relatively straightforward to accomplish. It was previously explained that a single sinusoidal signal with a non-Fourier-integer frequency will, in general, produce non-zero transform values at all bin numbers. Thus, with multiple sinusoids the transform rows in **R** are the linear superposition of the component **R** matrices associated with each sinusoid. Consider first the case where the time domain signal is comprised of two components with arbitrary and far separated frequencies (at bins m and n, n>>m) and comparable amplitudes. In this case the large separation between bins m and n results in the following behavior of the rows of **R**:

- the magnitude of the elements of **R** will be largest around the $m^{th}\pm1$ and $n^{th}\pm1$ rows (depending on the fractional ε values).
- the leakage effects will be small because of the large bin separation.
- each row of **R** will exhibit a beating pattern caused by the superposition of the two component **R** matrices.

The modulating envelopes at bins m and n will have relatively short periods due to the large bin separation, and the modulating amplitude will

be small compared to the mean amplitude. The accuracy of the estimates of the true transform magnitudes at these two bins (defined as the mean amplitude of at each row) will be reasonably unbiased because the segment length will typically be much longer than the short-period (and zero-mean) envelope modulations. The two frequencies and phases could be independently estimated directly from each transform row.

The idealized cases considered in this section were chosen to illustrate the properties and uses of this **R** matrix as clearly as possible. But sinusoids can be combined so many different ways that examination of further specialized cases is not productive. In the more general case the frequencies can be closely-spaced and the amplitudes non comparable such that the modulating effects in each row of **R** will be more complicated and the objective of identifying the initial frequency vector $\hat{\mathbf{f}}^{(0)}$ is more difficult. The remainder of this section describes a general methodology to estimate this vector for any multiharmonic signal.

Appendix A reviews some of the algebra of multiharmonic signals with emphasis on the beating signal resulting from the summation of two sinusoids. It explains that the effect of a third sinusoid is to superimpose a dynamic sinusoidal modulation to the usual single sinusoid envelope modulation. The methodology developed for estimating $\hat{\mathbf{f}}^{(0)}$ from an arbitrary **R** matrix builds on these two facts as follows:

- the length of the FFT used to generate (the rows of) **R** can be made arbitrarily long to allow for a bin resolution that isolates no more than three sinusoidal components.
- the number of time shifts used to generate (the columns of) **R** can be made arbitrarily long, subject to stationarity constraints, to insure that there are multiple modulation cycles in each row.

When these conditions are satisfied, it is assumed there will be at most three significant sinusoids contributing to each row of **R**; in many instances there will be only zero, one, or two significant sinusoids depending on the frequency distribution of the multiharmonic signal. The assumption is that for a large number of time shifts, the beating envelope corresponding to the two largest components can be successfully identified even from a true 3-component envelope.

The methodology proceeds as follows:

1. define $\mathbf{R}^{(0)}=\mathbf{R}$ and identify the row with the most energy.
2. estimate up to two sinusoidal components for that row:
   - 2.1 if the envelope is fairly constant, then fit *one* sinusoid (amplitude, frequency, phase)
   - 2.2 if the amplitude varies, then fit the best *two* sinusoids as follows:
     - 2.2.1 use the Hilbert Transform to calculate the envelope, and estimate its (difference) frequency
     - 2.2.2 estimate the mean (average) frequency

2.2.3 use Equation A.2 to algebraically estimate $f_1'^{(0)}$ and $f_2'^{(0)}$, i.e, trial approximations to the two frequencies; use these initial estimates to perform a numerical search (optimization) to find the best parameter estimates for the two-sinusoid fit to this row.

3. Append the two best estimated frequencies from the optimization to the estimate for the initial frequency vector $\hat{\mathbf{f}}^{(0)}$

4. Calculate component matrices $\mathbf{R}_1$ (and $\mathbf{R}_2$ if present) representing the transforms for these two identified sinusoids

5. Subtract the component matrices from $\mathbf{R}^{(0)}$ yielding an adjusted matrix $\mathbf{R}^{(1)}$

6. Identify the row in $\mathbf{R}^{(1)}$ with the largest transform magnitudes and repeat this adjustment process until:

   6.1 each row of $\mathbf{R}^{(N)}$ has negligible energy compared to the original $\mathbf{R}^{(0)}$ energy, or,

   6.2 further iterations do not reduce the energy in $\mathbf{R}^{(N)}$.

As stated in the assumptions section on the last page, this methodology can, in theory, be made arbitrarily accurate depending on the choices of the FFT length and the number of transforms (time shifts) used to generate $\mathbf{R}$. But these types of mathematically-based arguments that require very long data records are difficult to defend and apply for real world geophysical data such as the ocean waves emphasized in this application because of uncertainty over the stationarity characteristics. This dilemma is well known even in traditional spectral analysis. But keeping data segments

short to avoid possible temporal changes results in a loss of frequency resolution due the time-frequency ambiguity. Since different bandwidths of real world signals may, and probably do, have different time scales of stationarity (for example, low frequency swell versus the highest wind wave frequencies), then in any finite length data segment it may be unavoidable that nonstationarity is an issue at the high frequencies, while loss of resolution is an issue at the low frequencies. While it may be possible in some cases to low and high pass filter the data and perform independent analyses with different time steps and segment lengths, this is of questionable validity if the spectrum is not clearly bimodal (although further research is indicated).

With respect to this Harmonic Phase Tracking technique, errors are expected in $\hat{\mathbf{f}}^{(0)}$ for wideband signals because in practice selecting the FFT length and number of time shifts for **R** must be a compromise between maximizing low frequency resolution via long FFT lengths while simultaneously minimizing nonstationary effects at the higher frequencies via minimizing the number of time shifts. Of course, if all components are known to be stationary (as with some laboratory data) then both lengths can be made arbitrarily long.

Since the nonstationary characteristics of a signal are typically not known *a priori*, the general rule is to keep the segments as short as possible. This conservative approach may result in a wider than necessary frequency resolution from the FFT, with true, closely-spaced

components falling below the resolution limit and/or several components within one FFT bin. At such bins the proposed methodology to estimate $\hat{\mathbf{f}}^{(0)}$ can introduce errors, even for analytical signals defined as finite summations of [constant parameter] sinusoids. The source of this error is the present need to limit the model for each row of **R** to at most three significant sinusoids. This practice will estimate biased parameters if two sinusoids are within the resolution bandwidth or if three or more components are present with significant amplitudes. These component biases will in turn bias the adjustment process. However, since the subsequent Harmonic Phase Tracking technique adapts $\hat{\mathbf{f}}$ as necessary, any errors in this initial vector simply increase the number of iterations and are therefore not considered critical.

All of the methodologies presented in this Chapter require quantitative definitions and algorithms before they can be applied to measured data. Appendix C describes how this theoretical methodology was implemented in MATLAB, including descriptions and rationalizations for all the algorithms and threshold values.

Chapter 5 illustrates the performance of this new technique as applied to various types of analytical signals, harmonic and multiharmonic, with constant and non constant parameters, and with and without noise. Comparing the numerically-estimated and exact (known) parameters for these cases provides a firm basis for the analysis of laboratory and ocean waves in Chapters 6 and 7.

[blank]

CHAPTER 5

VALIDATION OF HARMONIC PHASE TRACKING USING ANALYTICAL SIGNALS

5.1 Chapter Overview

The motivation for the development of the Harmonic Phase Tracking (HPT) technique was the need for a more capable signal processing tool for the analysis of stochastic ocean wave fields in time and space. The mathematical and numerical phase of that development described in the last Chapter indicated that HPT indeed has valuable new capabilities worth investigating compared to existing techniques. Typically, the objectives of the next logical step in the development of any new technique such as HPT are to: (1) demonstrate accuracy using analytical and/or established experimental data, and (2) demonstrate robustness and uncertainty performance. Those characteristics are demonstrated and evaluated in this Chapter using analytical signals.

Normally, it is sufficient in the validation phase to present only a few examples that illustrate performance and applicability. However, ocean waves are very complex, with slowly-varying nonstationarities at unknown time scales that vary with frequency, and correlations to unknown length scales. So, before HPT results can be accepted with confidence for ocean waves, a very thorough validation step must be performed to establish how HPT models every one of these expected signal characteristics. This will satisfy the first objective regarding accuracy.

But that answer is incomplete; further fundamental questions still remain regarding robustness. For example, does the technique converge to one unique answer for all signal types, or does a change in the initial conditions or small perturbations to the data radically effect some estimates? And if so, under what circumstances, and by how much? And how does a discrete model such as HPT handle a signal with a possibly continuous spectrum? If any doubts remain on any of these questions, then the use of any new unproven technique like HPT on signals with unknown characteristics should be considered suspect. For those reasons, a wide-ranging validation study was undertaken using a variety of analytical signals with known characteristics.

These two comprehensive objectives make this is a long and important Chapter. A brief outline of the three sections that make up this Chapter is presented next.

Section 5.2 addresses the applicability of HPT to analytical signals with varying characteristics. The first numerical example uses a summation of 19 constant parameter sinusoids without noise to demonstrate how the technique handles a deterministic multiharmonic signal. Subsequent signals are analyzed in this section that exhibit constant and non constant parameters and additive white noise. The objective of Section 5.2 is to demonstrate that Harmonic Phase Tracking is capable of modeling all expected types of signals, as summarized in the following table:

Subsection	Signal Descriptor
5.2.2	summation of constant parameter sinusoids
5.2.3	single sinusoid with time dependent amplitude
5.2.4	one and two sinusoids with time dependent (chirping) frequency
5.2.5	bandwidth-limited white noise only
5.2.6	summation of constant parameter sinusoids with noise

Because the estimates from HPT converge asymptotically, they are therefore dependent on a variety of functions and convergence criteria (as reviewed briefly in Appendix C). So the next logical question to ask is: what is the numerical robustness and stability of the technique? This second objective is addressed in Section 5.3 using representative studies such as:

- affect of segment length versus frequency resolution (i.e, time-frequency ambiguity)
- affect of "errors" in the initial frequency vector estimate

Where appropriate, discussions are included regarding:

- accuracy of the estimated parameter set versus the true set,
- convergence characteristics, and
- comparison with traditional (FFT) spectral estimators using the same amount of information (i.e., comparable length of data vector).

The examples in these first two sections were selected to demonstrate the universality and applicability of the technique with as many types of signals as reasonable, given the length of this Chapter, rather than to exhaustively study the performance of HPT by concentrating on only a few signals. It was considered more important to establish credibility that HPT robustly handles <u>all</u> signal characteristics than it was to study optimum convergence strategies, minimum computation algorithms, or extensive numerical studies to identify error variance characteristics for each of the signal types. Results are presented graphically whenever possible.

Section 5.4 reviews the conclusions of this Chapter for analyzing deterministic and stochastic signals. This summary review then serves as the launching point for the analysis of laboratory waves in Chapter 6 and real ocean waves in Chapter 7.

5.2 Applicability of Harmonic Phase Tracking to Analytical Signals

5.2.1 Description of Multicomponent Analytical Signal

The first fundamental analytical studies were conducted using a multicomponent analytical signal comprised of 19 arbitrary constant parameter sinusoids as summarized in Table 5.1. While the rank and parameters were arbitrarily selected, the intent was to produce a signal with the following attributes:

- reasonably wide frequency band (note the ratio of approximately 5 between the highest and lowest periods)
- emphasize low bin numbers that have the least number of cycles in a given segment and are therefore the most difficult to identify
- vary the bin spacing between adjacent components to study the bin resolution (i.e., the time-frequency ambiguity) performance of the Harmonic Phase Tracking technique (e.g., the 7th and 8th components are very close, while the 16th is isolated)

No.	Bin Value	Period	Amplitude
1	4.6875	27.307	1.00
2	5.1875	24.675	1.00
3	5.5882	22.905	0.50
4	8.2500	15.515	0.50
5	8.8750	14.423	1.00
6	9.2353	13.860	0.50
7	9.8750	12.962	1.00
8	10.1250	12.642	0.50
9	10.7647	11.891	1.00
10	11.7500	10.894	1.00
11	12.1250	10.557	0.50
12	12.4118	10.313	1.00
13	12.7500	10.039	1.00
14	13.1875	9.706	0.50
15	13.7059	9.339	0.50
16	16.4375	7.787	1.00
17	21.6875	5.902	0.50
18	22.0625	5.802	1.00
19	22.7647	5.623	1.00

Table 5.1 Constant Parameters for Multicomponent Analytical Signal

Note to Table 5.1: The second column is defined relative to a 128-point segment length; integer values correspond to bin numbers for a 128-pt FFT. For example, the first sinusoidal component above has 4.6875 cycles every 128 points. The units for the third column are number of points (or, seconds if a unit time step is assumed for the time series) per cycle.

- vary the relative amplitude of adjacent components to study leakage effects.

The relative phases varied arbitrarily based on the starting index of the segment being analyzed so their initial values are not particularly relevant and are not listed here.

5.2.2 Representative Analysis of Multicomponent Signal

The objective of this first case is to illustrate the basic performance of the Harmonic Phase Tracking technique without complications due to noise or time-frequency ambiguities. This latter complication was avoided by purposely defining a segment length that was long enough to yield a resolvable bin width smaller than the known minimum bin spacing in the signal of 0.25 (between the 7th and 8th components). Using Equation C.5, a segment length of 360 points was selected, resulting in a "safe" minimum bin spacing of 0.177.

Figures 5.1 through 5.8 illustrate the results for this idealized stationary case. Since some of these same figures are used throughout this and subsequent Chapters, a full description is given for each, including what signal characteristics are best identified.

- Figure 5.1. Time domain representation of signals. The upper subfigure displays the 360-point segment (dotted line), with the final estimated time series superimposed (solid line). The lower subfigure displays the error between the original and fitted time histories as a solid line superimposed over the original segment (dotted line) for reference. It is evident from both subfigures that for this first case the Harmonic Phase Tracking technique accurately matched the time domain representation of the segment.

Note also that the total number of data points used for this fit is actually 540 points including the 90 points used for the forward and backward time shifts. The interpretation of how many data points are used for the HPT estimates is discussed further in Chapter 8.

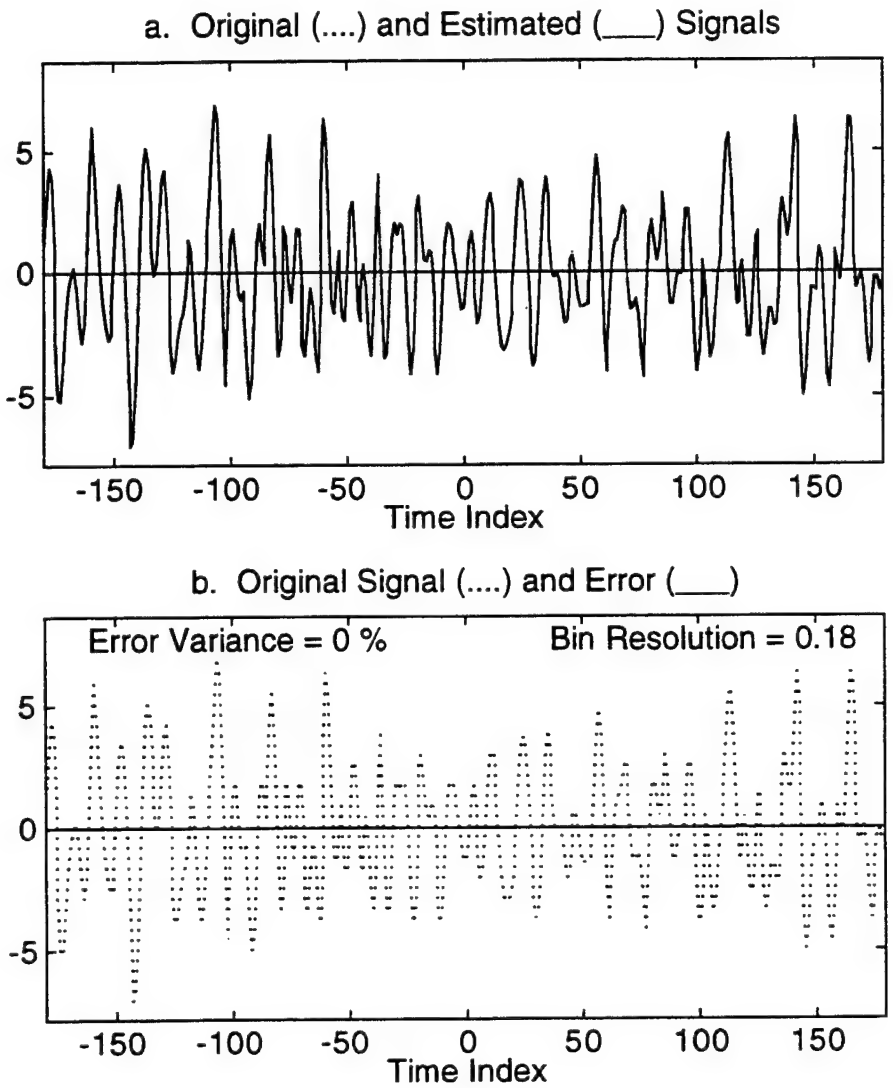

Figure 5.1 Time Domain Comparison of True and Estimated Multicomponent Analytical Signal

- Figure 5.2. Bin (Frequency) domain representation of signals. This figure compliments the time domain information in Figure 5.1 by displaying *amplitude* versus bin number (proportional to frequency). Note that this figure does *not* display energy as in the usual frequency domain plot of the spectrum. Consider a 2 component signal with sinusoidal amplitudes of 1.0 and 0.25. The amplitude of this second component is enough to make the time series envelope vary over a range from 0.75 to 1.25, and its presence would be readily recognized by inspection of the time series. On the other hand, the significance of this second component would be less apparent from the spectrum since its ordinate would be only $(0.25)^2$ or 6.25 percent as large as the main ordinate.

Therefore, this study uses plots of amplitude versus frequency instead of energy versus frequency as more representative of the relative significance of component amplitudes.

Exact, HPT-estimated, and FFT-estimated amplitudes are superimposed in Figure 5.2. The FFT amplitude is defined as:

$$\mathbf{c}(f_j \,|N_{FFT}) = \left(2/N_{FFT}\right)\left|X(f_j)\right| \qquad 5.1$$

where: **c**=discrete amplitude vector which is a function of the FFT length N_{FFT} (here, 512 points was used as the closest power-of-2 length to the 540 points used for the HPT estimates), $f_j=j^{th}$ discrete Fourier

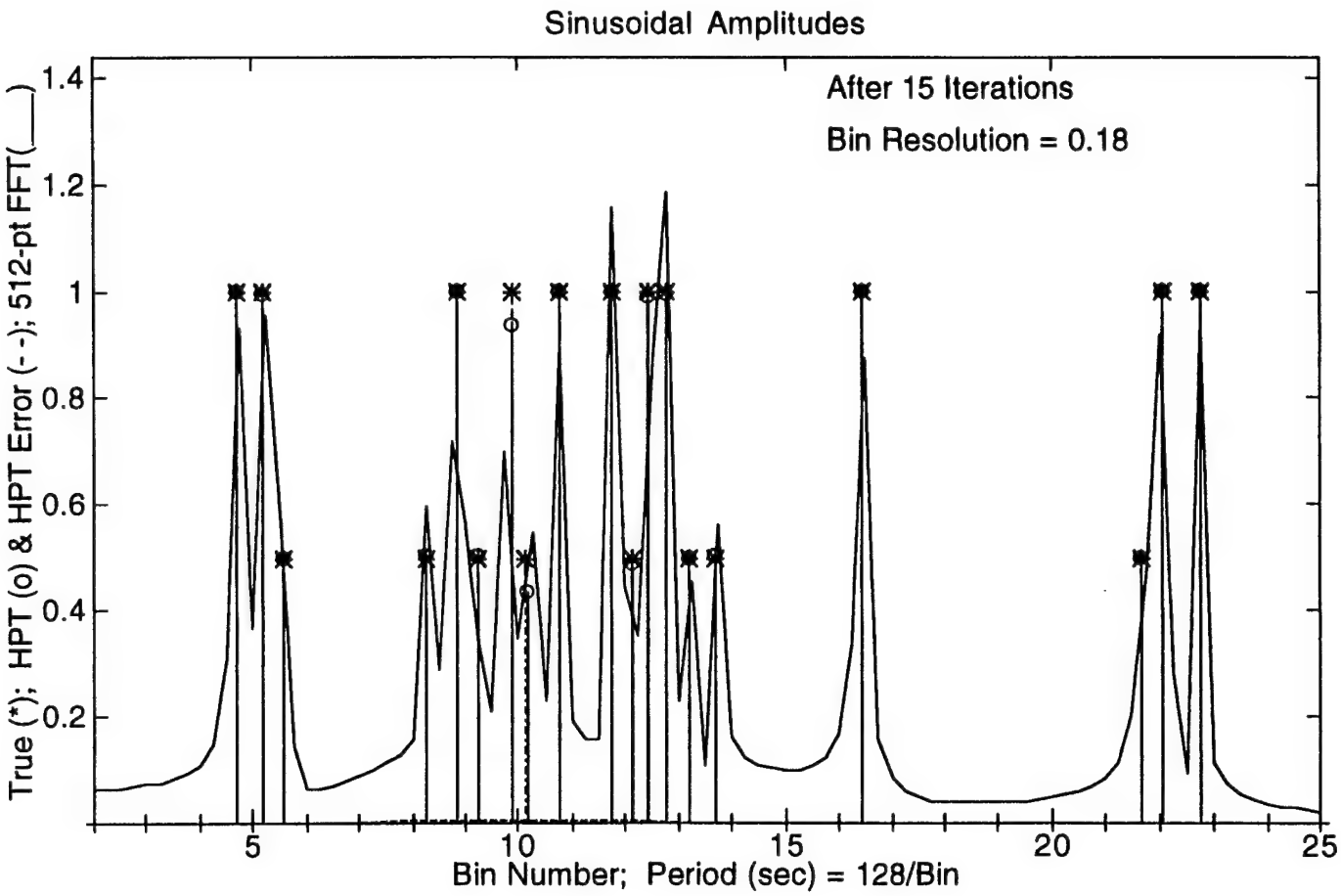

Figure 5.2 Bin (Frequency) Domain Comparison of True (*) and Estimated (o) Multicomponent Analytical Signal

frequency, and $|X(f_j)|$=FFT ordinate of the signal. A simple rectangular window was used for this spectral function, and unit time steps are assumed with no loss of generality. A similarly-defined FFT amplitude function of the error between the original and HPT-fitted time series is also plotted as a dashed line.

Figure 5.1 showed that HPT accurately fitted this time domain signal. That is, of course, a desirable condition for a good estimator. But Figure 5.2 illustrates the key feature of this new technique - it demonstrates that *Harmonic Phase Tracking is capable of accurately identifying the true set of components (rank and frequency values).* It did not converge to an arbitrary set of [orthogonal] components with arbitrary rank (the category that Fourier Series fits in) just to fit the time domain signal. This is the first demonstration of this unique and powerful feature of the Harmonic Phase Tracking technique, and this ability will be the central focus of the remaining comparisons in this and subsequent Chapters. The estimated and true frequencies and amplitudes for this case are shown in the Table 5.2.

It is informative here to briefly discuss under what conditions a traditional FFT-based spectrum such as the one defined in Equation 5.1 would identify these 19 components. First, it would require a minimum FFT length of 512 points to reduce the resolvability to the necessary 0.25 bin spacing (or, practically speaking, more than 512 points to

Estimated Bin Number	True Bin Number	Estimated Amplitude	True Amplitude
4.687	4.688	1.000	1.000
5.187	5.188	0.998	1.000
5.590	5.588	0.498	0.500
8.250	8.250	0.501	0.500
8.875	8.875	1.002	1.000
9.235	9.235	0.500	0.500
9.862	9.875	0.936	1.000
10.150	10.125	0.435	0.500
10.765	10.765	0.999	1.000
11.749	11.750	1.000	1.000
12.121	12.125	0.488	0.500
12.414	12.412	0.990	1.000
12.751	12.750	0.997	1.000
13.187	13.188	0.500	0.500
13.706	13.706	0.502	0.500
16.438	16.438	1.000	1.000
21.687	21.688	0.500	0.500
22.063	22.062	1.001	1.000
22.765	22.765	0.999	1.000

Table 5.2 HPT-Estimated and True Parameters for Multicomponent Analytical Signal

slightly exceed that resolution to minimize leakage). That minimum number of points seems to correspond favorably with the 540 used by the harmonic phase tracking technique. But even that FFT length would only insure that one sinusoid was contained somewhere within a bin - however, *the exact frequency, amplitude and phase would still be unknown*. A second fundamental flaw comes in the next step, specifically, that it is necessary to calculate several transforms over independent segments in order to perform ensemble averaging. This is unavoidable when the signal properties are unknown. Chapter 8 presents more information on uncertainties for these and other estimators.

- Figure 5.3. Reduction of the Error Spectral Function. This figure displays how the [FFT-based] spectrum of the error was decreased versus HPT iteration. The ordinates are normalized by the peak ordinate of the FFT amplitude function of the signal. The orientation is such that the final error function is the front row. Note how the error decreases for each bin number as the iterations progress. As detailed in Appendix C, the final criteria for convergence are subjective; in this case the small residual signal was considered asymptotically converged and not significant. This residual error function can be useful in verifying whether this component signal (in essence, "noise") matches the assumptions made in the modeling; it can also be informative for showing which frequencies are difficult to fit, which

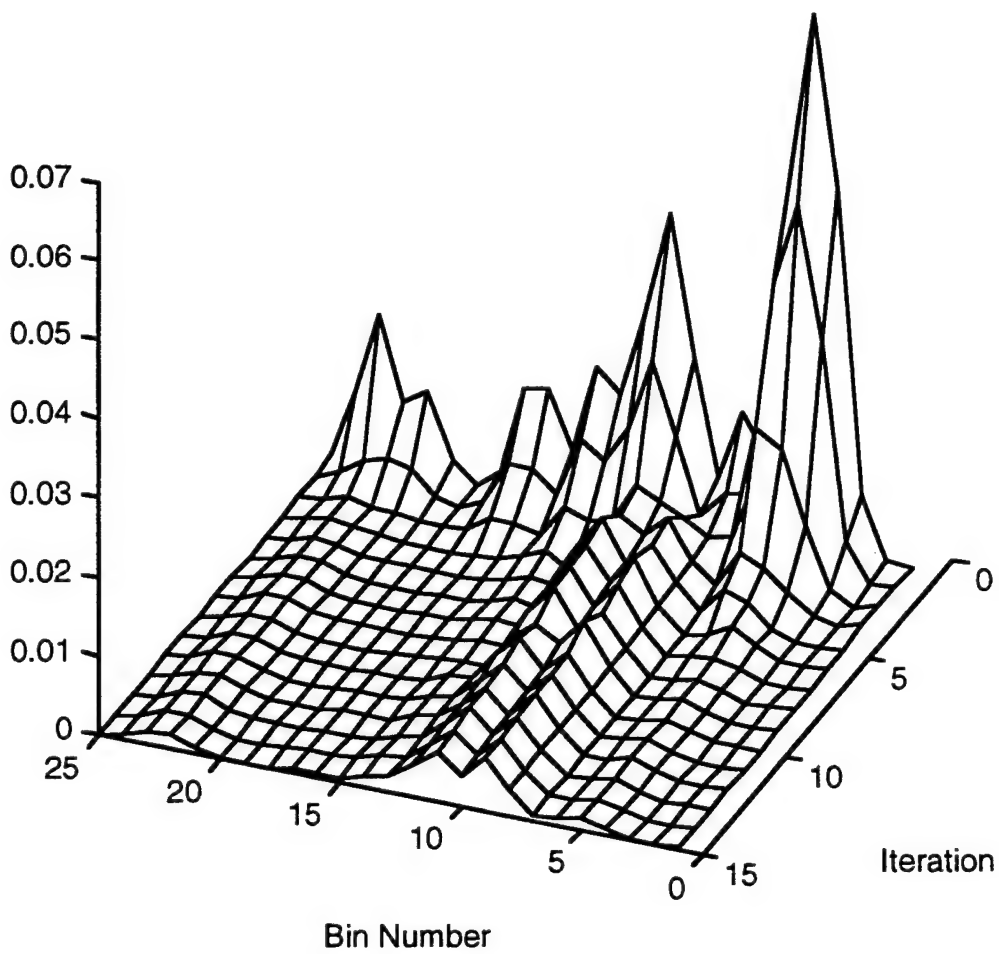

Figure 5.3 Distribution of Normalized Residual Error versus Bin Number and Iteration Number for Multicomponent Analytical Signal

implies that there are no well-formed sinusoids in that frequency region. Note also that this HPT analysis required 15 iterations.

- Figure 5.4. Evolution of the RMS Error(Time Shift) versus Iteration.

The rms error versus the forward/backward time shift is calculated as follows: start with the initial segment of the time history and the initial frequency vector; fit amplitudes and phases; and calculate a root-mean-square error σ_0. Shift the time series forward and backward and recalculate a σ_j for each segment. Plot σ_j versus the time shift as the furthest-back row in this figure (i.e., first iteration). Then, after modifying the frequency vector, repeat and replot the new σ_j vector as the second row. Repeat for each iteration (here, 15). Thus, the row in the forefront represents the error in the final fit as the time series is shifted forwards and backwards.

This final row can be interpreted as a scalar measure of the stationarity of the frequency vector, but not necessarily of the signal because the amplitudes and phases are independently fitted for each shifted segment; those functions can be inspected for stationarity as discussed in the text for the next figures.

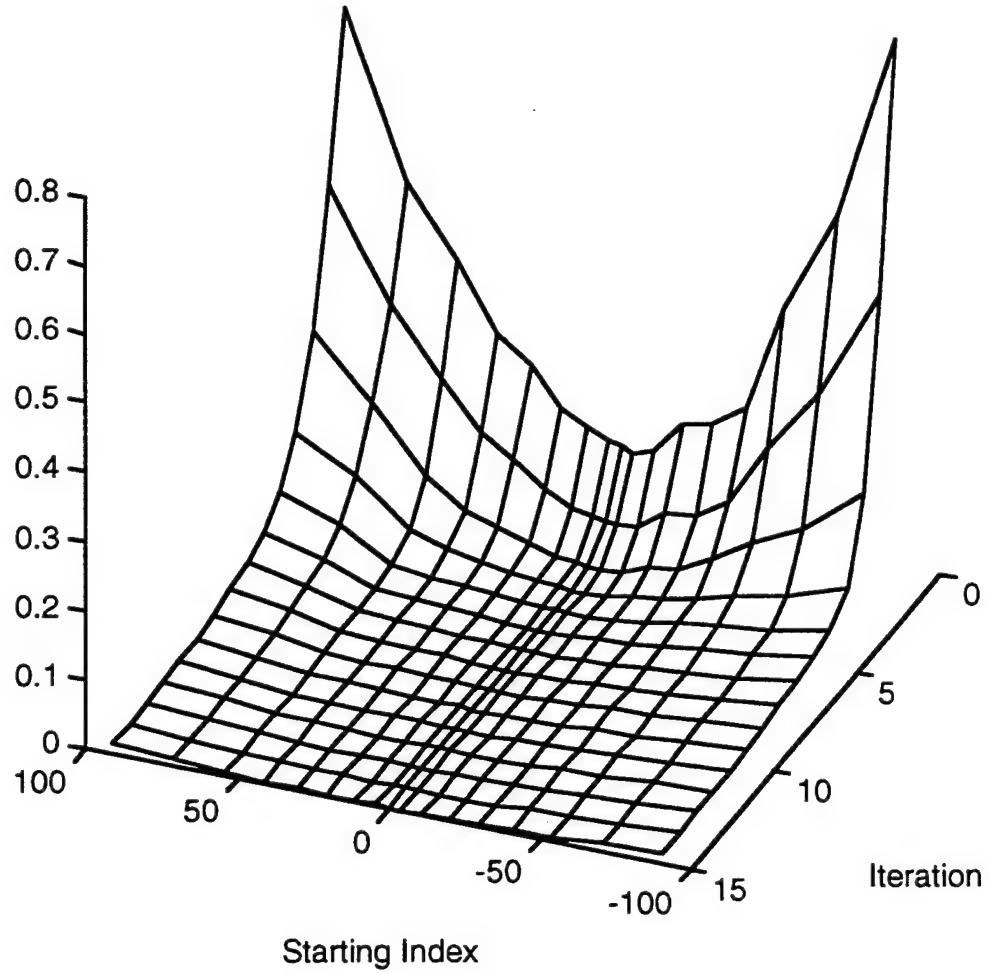

Figure 5.4 Error in Fitting Each Shifted Segment versus Iteration Number for Multicomponent Analytical Signal

- Figure 5.5. HPT-based Amplitudes versus Time Shift. This is a plot of amplitude versus frequency (bin number) as the segment is shifted. Since the frequency vector is constant for this set of shifted HPT results, then constant amplitudes on that figure would strongly indicate a high degree of stationarity in the signal components. In fact, the information available in these last two figures could provide new quantitative definitions for various vector measures of stationarity that may be superior to presently-used scalar and moment measures for multiharmonic signals (discussed in the last Chapter).

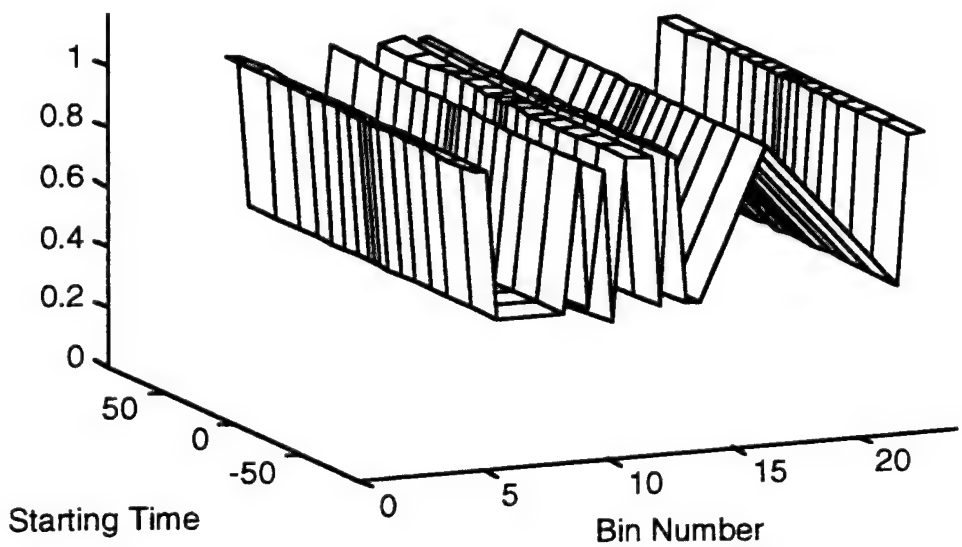

Figure 5.5 Stationarity of Amplitudes versus Starting Time for Multicomponent Analytical Signal

- Figure 5.6. Sample Parameter Convergence versus Iteration. This figure illustrates convergence of one representative component during this fitting process. The three subfigures show how the frequency, amplitude, and phase converged versus iteration number; see Table 5.2 for the true and asymptotic numerical values.

- Figure 5.7. Signal Extrapolation Using HPT Parameters. While the Harmonic Phase Tracking technique does require forward and backward time shifts to identify the frequency vector, the "reference" amplitude and phase vectors at each iteration are calculated using only the center segment of the time history (here, the 360 point segment). It is suggested that using those "center" amplitudes and phases to extrapolate the signal beyond the center segment would serve as an alternative test of accuracy and stationarity for the estimated parameter set. The purpose of this figure is to illustrate the concept. The upper subfigure shows backwards extrapolation (to the left of the dashed line at the zero time index) while the lower subfigure shows forward extrapolation (to the right of the dashed line at 360). The original data is the solid line while the fitted/extrapolated data is the dotted line (masked here by the solid line). An alternative extrapolation technique based solely on stationarity of the frequency vector is discussed in the final Chapter.

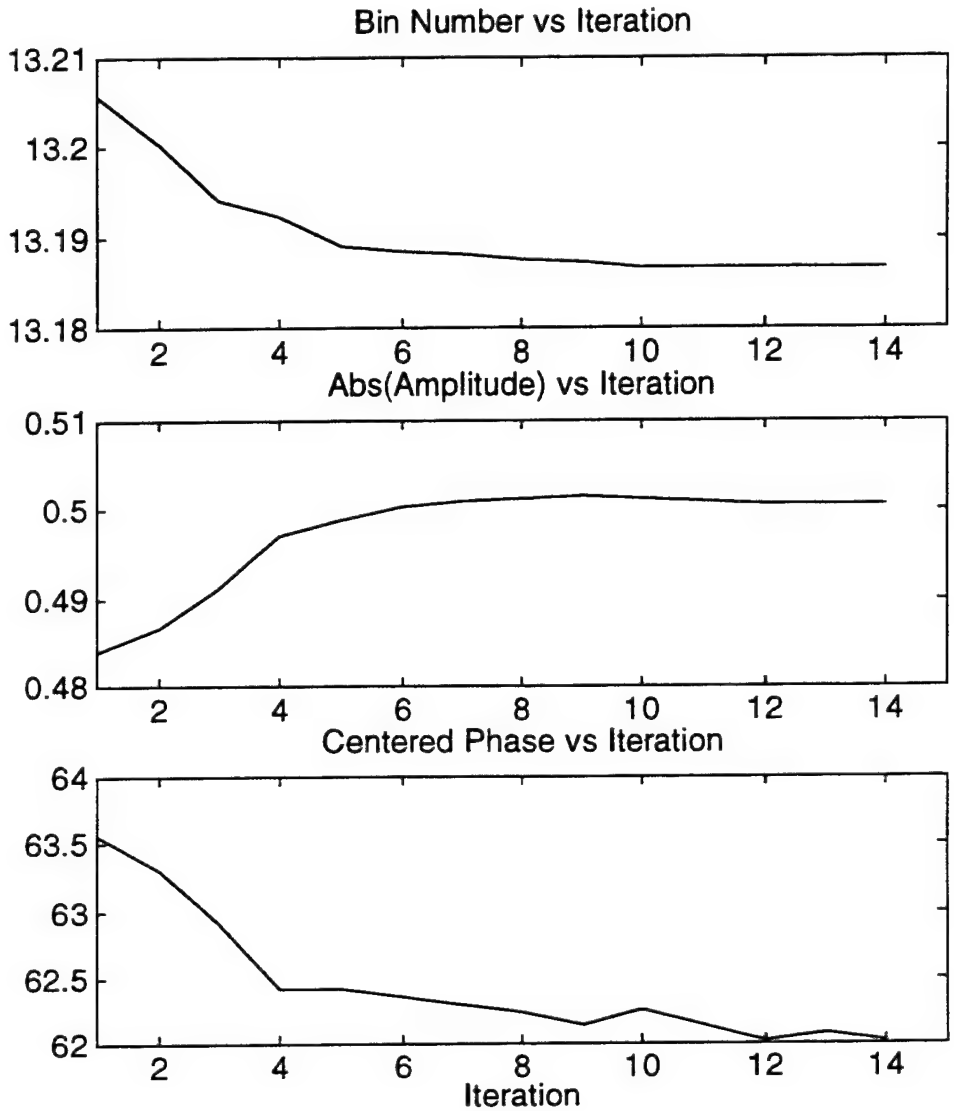

Figure 5.6 Sample Convergence of HPT Parameters for Multiharmonic Analytical Signal

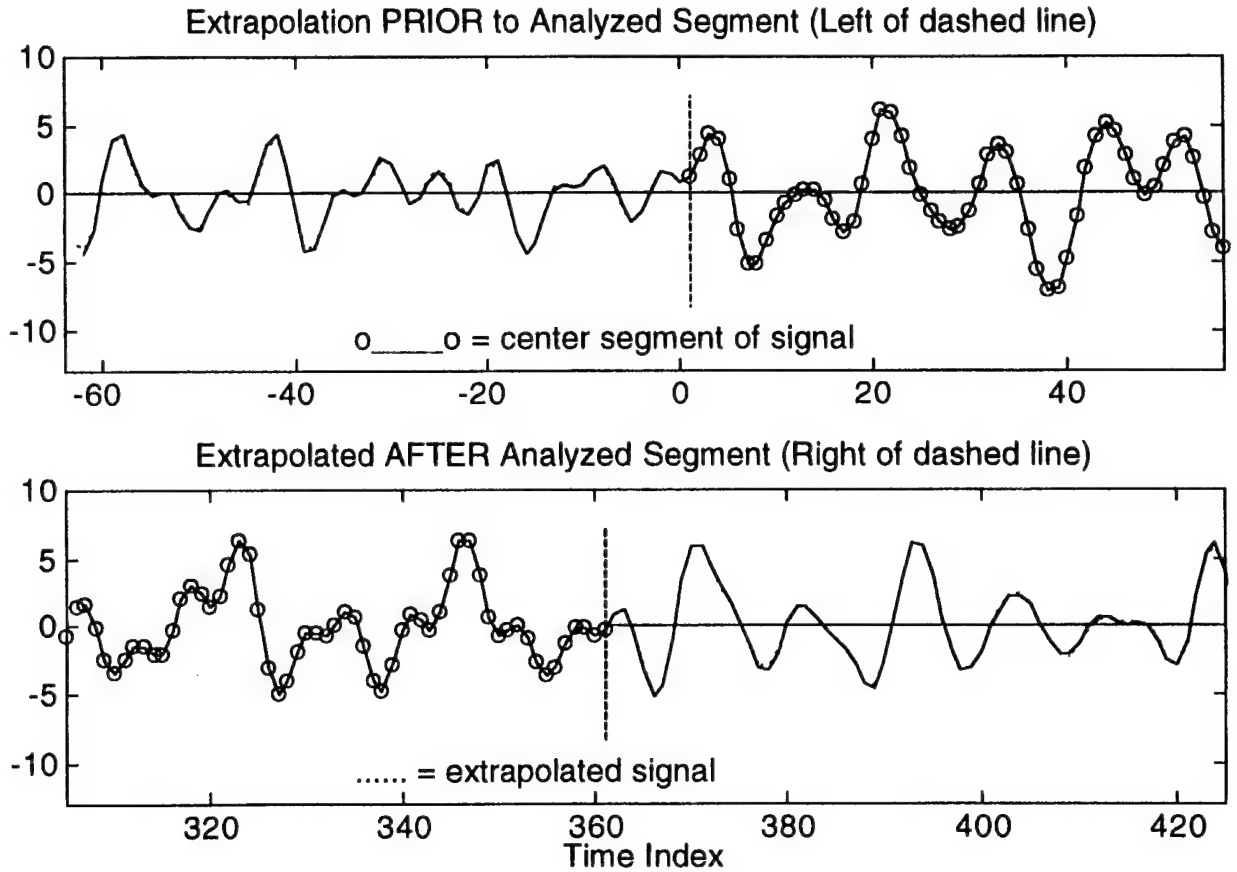

Figure 5.7 Example of Signal Extrapolation Using HPT Estimates

Regarding Figure 5.7, note that extrapolation is undefined for Fourier Series representations, which reinforces the fact that Fourier coefficients should not be confused with physically-present harmonic components.

In summary, this first set of figures for the multiharmonic signal shows that the Harmonic Phase Tracking technique is robust, accurate, and stable when properly applied to a stationary, deterministic signal comprised of a finite summation of constant parameter sinusoids. That is an important first conclusion to establish. This analysis using uncompiled code in MATLAB 4 required under 10 minutes on a Sun SPARC10 unix computer. The condition numbers of the real and imaginary basis matrices when converged were only 3.0, indicating that the high correlation (due to the close frequency separation) between some of the components did not significantly affect the results. Performance of HPT when there are components closer than the resolution limit is reviewed in Section 5.3.

5.2.3 Representative Analysis of Single Sinusoid with Time Dependent Amplitude

The previous analysis searched for a multicomponent deterministic signal defined using constant parameter sinusoids. This and the next subsection investigate the effect of temporal (nonstationary) perturbations in the frequency or amplitude.

HPT returns a very interesting estimate for a single sinusoid with a linearly-varying amplitude (but fixed frequency and phase). Any segment of this signal can be modeled using a local, zero-mean time t' as

$$\begin{aligned} x(t') &= \left(a_o + a_1 t'\right) \sin\left(2\pi f t' + \theta\right) \\ &= a_o\left(1 + m\ t'\right) \sin\left(2\pi f t' + \theta\right) \qquad -T \le t' \le T \end{aligned} \tag{5.4}$$

where $m \equiv a_1/a_0$. This is a very natural way to model the signal since a_o is simply the mean amplitude over the segment that a parameter estimation technique such as HPT, MUSIC, etc. would find; Fourier Series could find it as well but only in the special case when the frequency corresponded to an integer Fourier harmonic. Define a_o=1 for simplicity and rewrite Equation 5.4 as the sum of stationary and nonstationary parts:

$$x(t') = \sin\left(2\pi f t' + \theta\right) + \left[m\ t'\right] \sin\left(2\pi f t' + \theta\right) \qquad -T \le t' \le T \tag{5.5}$$

The question is, how does the nonstationary term affect the estimated parameters and rank? The answer is that *Harmonic Phase Tracking is unaffected by the presence of this second term, regardless of the relative values of* a_o *and* $a_1 T$. While this seems like an unexpected response, it can be justified using two independent arguments.

For illustrative reasons, first consider the case when the phase is zero. The nonstationary function is then $t' \sin\left(2\pi f t'\right)$, which is an even function. This is straightforward to visualize, since the basic sinusoid must be

increased in phase for positive t', yet decreased (out of phase) for negative t'; thus this "correction" term has a 180 degree phase discontinuity at the origin. There is no *single* frequency sinusoid that can model this function because of the discontinuity at zero. This argument holds equally for cosinusoidal signals also.

Secondly, this $t'\sin(2\pi f t')$ function, with the 180 degree phase discontinuity at the center time, looks exactly like a beating signal produced by two equal amplitude sinusoids as reviewed in Appendix A. So it is natural to next ask if HPT attempts to model this correction term using *two* instead of one sinusoid. Here there are two answers. If the $t'\sin(2\pi f t')$ function <u>alone</u> is fitted, then yes, the technique will identify two components described by the following:

- equal amplitudes, $|a_1| = |a_2|$
- 180 degree difference in phases; the envelope is zero for local time t'=0, requiring $x_1(0) + x_2(0) = 0$, or, $\sin(\theta_1) + \sin(\theta_2) = 0$ which in turn requires $\theta_2 = \theta_1 + \pi$.
- but most crucially, the envelope period is very long, requiring that the two component frequencies be very close. Strictly speaking, this spacing is undefined, because the linear variation in the envelope corresponds to an infinite number of envelope periods and envelope amplitudes (using the expansion $a\sin(2\pi f t) \approx a\left[(2\pi f t) + O(t^3)\right] \approx (2\pi a f)t$).

Harmonic Phase Tracking avoids this ambiguity problem and finds a two component best-fit because the solution is based on *total* rather than standard least squares. Incorporating the right hand side vector into the basis matrix pulls the "projection vector" out of the range of the usual basis matrix (where there is no solution). As a result, the HPT solution converges to two (equal) component amplitudes whose sum by definition becomes the maximum amplitude of the envelope. With the known linear slope and the amplitudes arbitrarily quantified this way, the technique can resolve the ambiguous amplitude-frequency product and subsequently converge to a set of component frequencies whose bin spacing is consistent with these amplitudes and whose average frequency is the true signal frequency. Also, the phases will show a 180 degree difference as required. This is a good illustration of how the use of total least squares greatly increases the stability of the technique for real-world signals (i.e., it decreases the variance by introducing some bias).

Now return to the issue of how HPT analyzes the complete signal defined by the stationary and linearly-varying components in Equation 5.4. For simplicity assume $a_0 > a_1 T$ (to avoid an awkward phase discontinuity in the signal when the amplitude switches sign). In these cases HPT will only estimate one component with the (correct) mean amplitude a_o at the true frequency. The explanation for this is that HPT finds the predominant component and frequency (the a_o component) first. It then tries to model the 2 beating components, but the estimated frequencies are often too

close to the existing mean frequency and they are therefore not inserted by the technique. The fact that HPT models or does not model the nonstationary component of the signal depending on whether there is a second component present further illustrates why the technique is not strictly a "linear operator."

Is the fact that the technique does not model such a nonstationary signal component a weakness? If the sole measure applied is the rms error in the time domain then, yes, this is a weakness. But it has been stated that a unique feature of this technique is its ability to identify the true number of sinusoids (i.e., the rank) in a signal, and by this measure this inability to add extraneous components to model the change in amplitude is in fact a strong advantage. Viewed another way, this technique is complimentary to techniques such as Fourier Series, which has the advantage of accurately modeling any physically-realizable time domain signal but with the corresponding disadvantage of doing so with a set of components that very rarely have any physical meaning. The final decision as to the superiority of one technique over another cannot be made here - it depends on the objectives of the analyst and the problem at hand. But it can be said that by offering a different set of features the Harmonic Phase Tracking technique has already established its value.

The preceding discussions have focused on linear variations in amplitude. If higher order amplitude variations such as quadratic and cubic

variations are present, this minimizes the amplitude-frequency ambiguity and Harmonic Phase Tracking interprets the nonstationary signal as a pair of beating sinusoids rather than one predominant sinusoid, and subsequently it does result in an approximate fit to the nonstationary signal (using the time domain error measure).

The fact that the technique interprets changes in amplitude differently depending on whether they are linear or not does introduce some complications to analyzing signals with unknown properties. Consider a signal defined as a pair of closely-spaced, unit amplitude sinusoids such that the length of data selected for the analysis window is approximately 60 degrees of the envelope cycle. (Recognize immediately that for most applications this would be an extraordinarily-short segment, representing only 1/6 of a cycle for that signal; put another way, it is 6 times shorter than the length corresponding to a comparable bin resolution of Fourier Series.) If the segment is centered across an envelope crest, then the signal looks like *one* sinusoid with a very small change in amplitude; HPT will estimate one sinusoid, with a mean amplitude of approximately 1.8, at the mean [beating] frequency. Shift the segment forward to where the envelope appears like a linearly-decreasing amplitude between the crest and a node; now the technique estimates *one* sinusoid, at the same mean frequency, but with an amplitude of 1.0 (since it does not see the essentially linear amplitude variation). Shift the segment again to straddle a node; now the technique will estimate *two* components,

approximately equal to the two true frequencies, but with a slightly biased set of amplitudes (slightly lower than one depending on the length of the data segment).

Thus, three analyses of the same signal produced three *qualitatively* different estimates. While this is not particularly desirable behavior, recall that it happens only when the length of the analysis segment is well below normal practice. However, since there is never a guarantee that this type of very low frequency component will not be present in a finite segment of a signal (unless careful low pass filtering is applied to the longer record), it is valuable to understand how it is modeled and how it may affect other neighboring component estimates. If multiple analyses are done on sequential segments of the data, there would be an observable oscillation in the frequency domain around one dominant frequency that would allow for confident identification of the fact that the signal contained two sinusoids. This might not be as evident if a large level of noise was present, so this ensemble inspection process may not be as straightforward as it seems for interpreting real-world signals. Observing such inconsistencies in the frequency domain results would strongly indicate that there was a potential resolution ambiguity around that frequency region. The solution would be to repeat the analyses with a longer test segment to reduce the minimum bin spacing to decrease the resolution. These results are further examined in Subsection 5.3.1.

The Fourier Series representation of a segment of this type of nonstationary signal is best understood from inspection of the time series defined in Equation 5.5. If the fundamental frequency and segment length do not correspond to a Fourier harmonic, then the first term in Equation 5.5 will itself produce finite ordinates at all the Fourier bin numbers. Second, the second term in that equation (illustrated in Figure 5.16) is certainly nonsinusoidal so that the transform will require ordinates at all the Fourier bins. So, a single Fourier transform/spectrum of the total signal would be the sum of these two "wideband" transforms, and accordingly it would be very difficult to interpret.

The Harmonic Phase Tracking technique required 15 to 35 iterations for the types of signals and segment lengths used in these studies.

5.2.4 Representative Analysis of Sinusoids with Time Dependent Frequency

The second type of nonstationarity that would violate the assumption that the signal is a [summation of] constant parameter sinusoid occurs when the frequency (or, equivalently, the phase) varies with time. Accordingly, signals with linear and oscillatory frequency variations (but constant amplitude and phase) are investigated in this subsection.

The fundamental signal over any time interval of length 2T is defined for a linear frequency variation by:

$$x(t') = \sin\left(2\pi\left[1 + mt'\right] ft' + \theta\right) \qquad -T \le t' \le T \qquad 5.6$$

where a unit amplitude is used with no loss of generality.

In the last subsection it was shown that the Harmonic Phase Tracking technique either did not model an amplitude variation (for linear variations), or, it naturally approximated the change using two constant parameter sinusoids. In contrast to that nonstationarity, modeling this continual variation in frequency is numerically more difficult. Just as with the modeling of nonstationary amplitude, the technique uses combinations of constant parameter sinusoids to approximate the varying frequency.

Conclusions here are not as direct as with the former case. Some observations are possible:

- if the frequency variations are small (less than 3 percent), then the technique estimates only one component at the mean frequency for that segment. The reason is that the error term between the original and this rank-one fit signal looks approximately like a beating signal with a node at the center, and the technique follows the same behavior as described in the last subsection for amplitude nonstationarity.

- unlike the situation for amplitude changes, modeling larger frequency variations starts with a rank-1 sinusoid but adds other sinusoids at neighboring frequencies. For example, for a single, unit amplitude sinusoid with a five percent frequency variation, the Harmonic Phase Tracking technique estimated the following rank-3 fit:

Bin Number	Period	Amplitude	Phase (deg)
4.334	29.54	0.113	3
4.600	27.83	0.991	91
4.872	26.27	0.114	-5

Note how the "sideband" components have equal amplitudes and phases associated with an even, beating signal. Other studies with larger frequency variations resulted in even higher rank approximations. The following table shows a representative fit from one of those studies that used a 180 point single unit amplitude sinusoid at bin 10 (relative to a 64-point FFT) with a 10 percent frequency change over 300 points:

Bin Number	Period	Amplitude	Phase (deg)
9.37	6.83	0.280	85
9.57	6.69	0.516	-129
9.91	6.46	0.612	-48
10.31	6.21	0.616	-73
10.48	6.11	0.401	156

This rank-5 estimate is qualitatively quite different from the original rank-1 nonstationary sinusoid, yet the good time domain fit shown in Figure 5.8 attests to its absolute accuracy. The instantaneous frequency was calculated using Equations A.12 through A.14 for both the original and the estimated signals and the agreement was excellent. Both of these measures confirm that the HPT fit is accurate, at least within the chosen segment. Note that accurate extrapolation is not reliable using constant parameter harmonics when the signal is nonstationary as in this example.

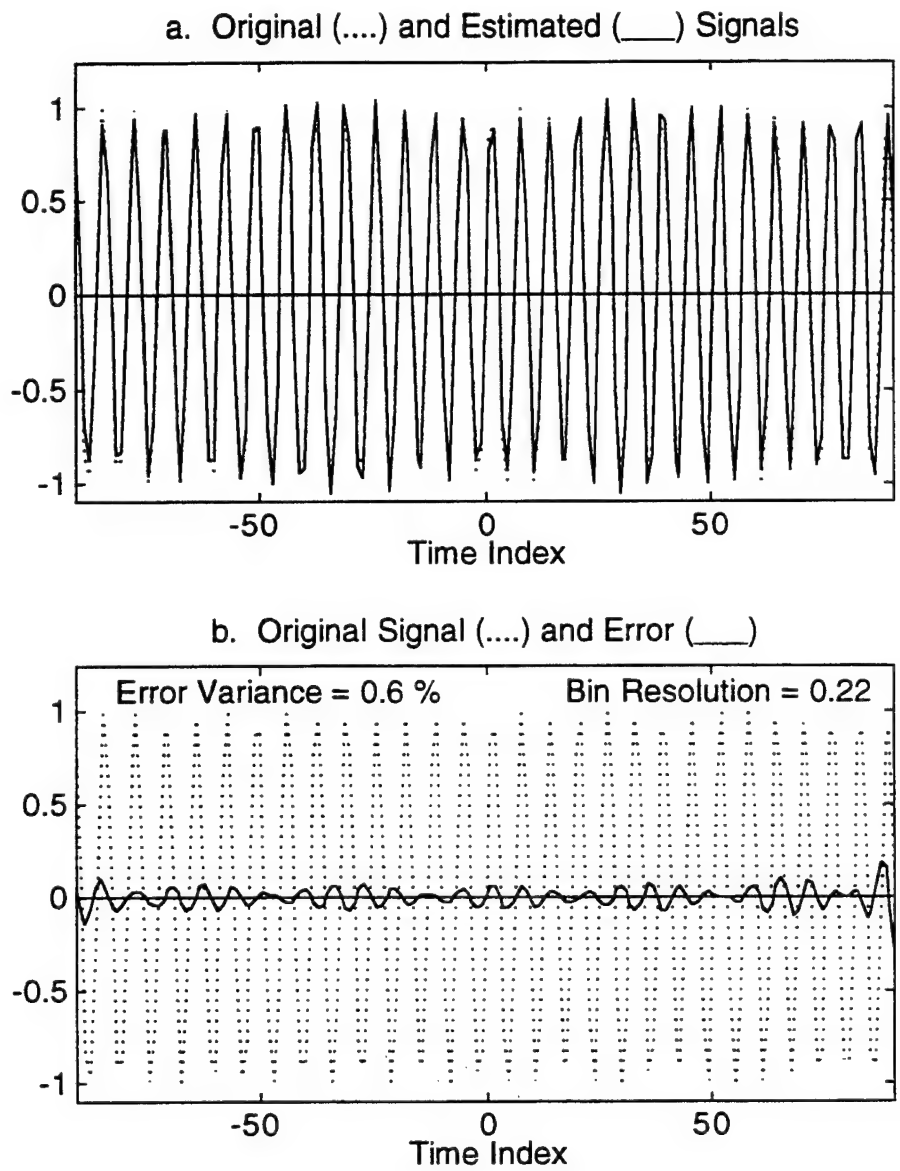

Figure 5.8 HPT Estimated Components versus Bin Number for Chirped Sinusoid with 10% Frequency Nonstationarity

But what is the geometrical basis for this particular combination of five constant parameter sinusoids found using Harmonic Phase Tracking? The answer can be found the subtle beating behavior of the HPT estimate evident in Figures 5.8. As described in Appendix A, recall that instantaneous frequency is a function of the "local" energy of the signal components (where "local" accounts for the time dependent term in Equation A.15). Now imagine a signal comprised of four sinusoids where the first pair makes a beating signal such that the envelope starts at a crest and ends at a node over the segment being analyzed, and the second pair makes a beating signal with a mirror image envelope (i.e., crest at the end). At the beginning of the record the first pair would dominate the instantaneous frequency calculation, while the second pair would dominate at the end of the segment. Last, add a fifth sinusoid that modulates both envelopes such that there is an additional envelope crest in the center of the segment; this fifth sinusoid helps keep the amplitude close to constant while also acting to smooth the transition between the two pairs.

It was concluded from this and the last subsection that HPT is more sensitive to frequency rather than amplitude variations. It was therefore considered prudent to better understand how HPT models signals with frequency nonstationarity. So, two more examples of signals with nonstationary frequencies are presented, and a new graphical display is introduced to illustrate the results.

The first signal has one sinusoid with a linearly varying frequency like the last example. However, a new type of study and display is introduced whereby a series of HPT estimates are made on overlapping segments. After the first estimate is completed, the start time for the segment is shifted forward - in this case by 50 percent of the segment length - and a second analysis is performed. After repeated analyses the results can be graphed similar to a traditional "waterfall" plot to display variations in frequency and amplitude. 16 total analyses were done using 200 point segments.

Results for this first signal are shown starting with Figure 5.9. The middle subfigure shows the signal versus time for reference, with a wide vertical bar near the center indicating the HPT segment length of 200 points (just over 3 minutes on this plot assuming a unit time step). The right subfigure is, in essence, a version of the FFT-based "waterfall" diagram. The diameter of the circles represent the amplitude (not energy) of each Fourier Series fit; open circles represent estimates and the filled circle is the true (unit) amplitude. Bin numbers on both the left and right subfigures are arbitrarily referenced to a 128-point FFT; since the HPT analysis used a 200 point segment, a comparable FFT length of 256 was chosen. Thus, the frequency resolution for the right hand subfigure is half bin numbers because the FFT length of 256 points is twice as long as

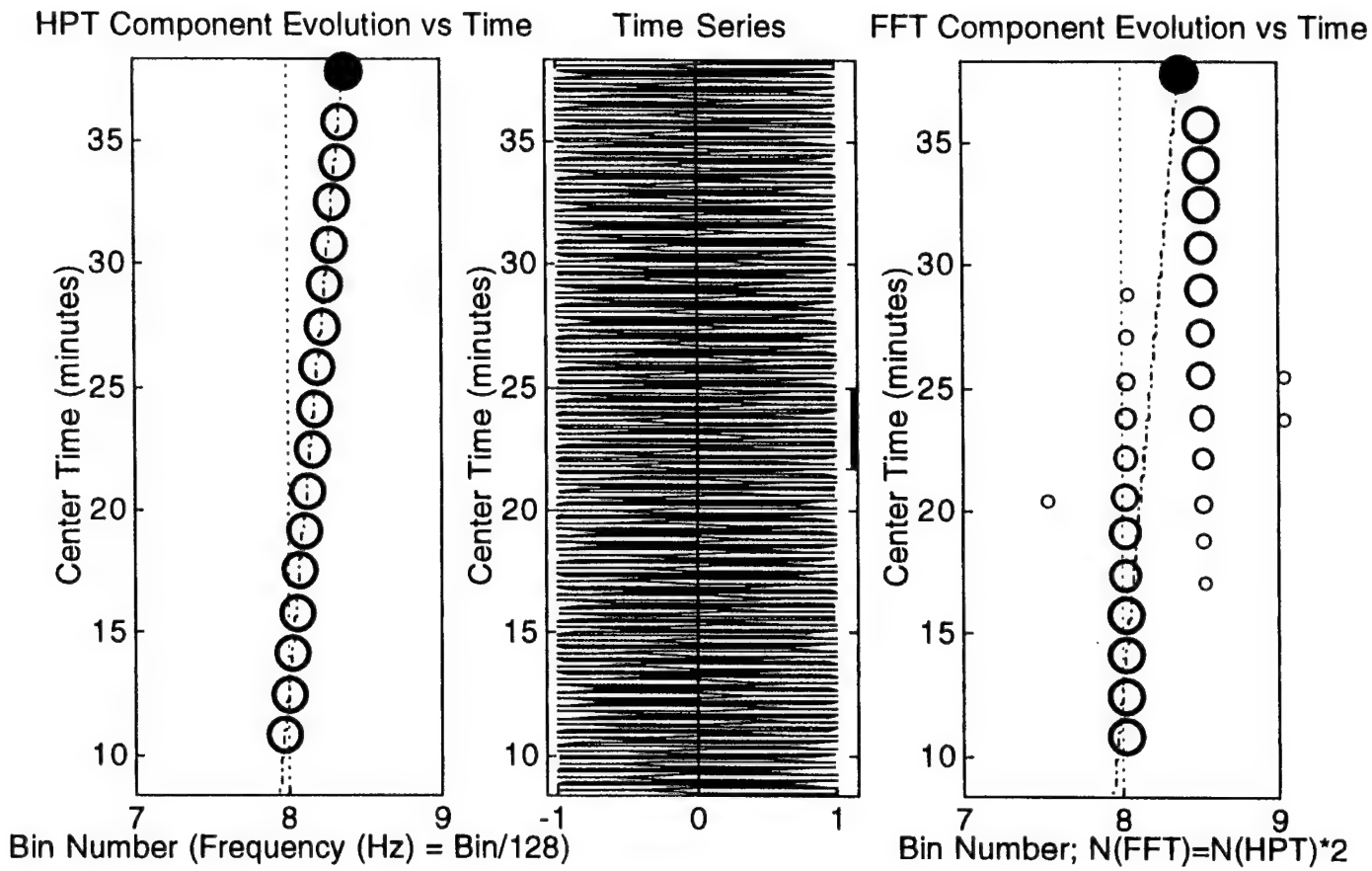

Figure 5.9 HPT- and FFT- Estimated Component Evolution versus Bin Number for Sinusoid with Linear Frequency Nonstationarity

the "reference" 128 points. The dashed line is the numerically-derived instantaneous frequency.

While the FFT ordinates do show a steady shift to higher bin numbers, it is not possible from the FFT plot to identify the rank or any other parameters of the signal(s).

The left subfigure in Figure 5.9 shows the HPT results. Again, the circle diameters represent amplitude, with the same scaling and meaning as the right subfigure. Note that the HPT results are much more revealing than the FFT results, since HPT correctly identified one sinusoid with a varying mean frequency. Subsequently, the amplitude estimates are correct also.

The low rank, "smoothness" of the HPT frequency evolution versus time, and invariance of the HPT amplitude estimates in Figure 5.9 are all persuasive indicators that HPT has accurately fitted the signal characteristics, especially in cases where the characteristics weren't *a priori* known. Figure 5.10a, which is an alternative display of the left subplot in Figure 5.9, does indeed show how precise the HPT estimates really are for this example signal. All of these observations make the HPT results more informative and/or more believable than the FFT results. But what is really needed is a function that would provide an analyst with a *quantitative* measure of whether a particular a continuous set of HPT estimates truly represented a coherent sinusoid evolving (or not) over time. A valuable feature of HPT is that it does provide such a measure.

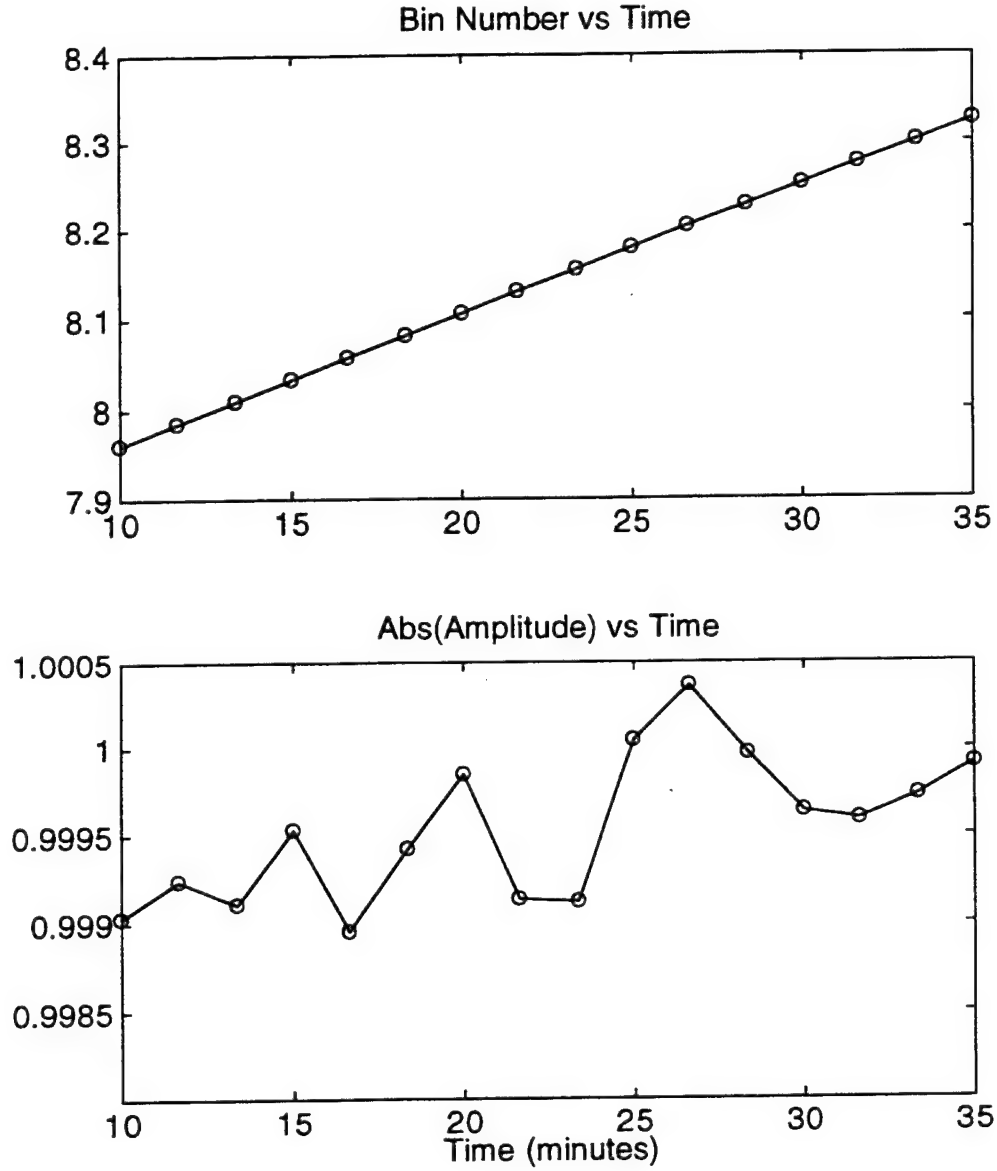

Figure 10a. HPT-estimated Bin Numbers and Amplitudes versus Time for a Signal with a Linearly-Varying Frequency

Figure 10b complements Figure 10a by showing HPT-estimated continuous phases versus time. Each circle represents the HPT estimated phase relative to the center of a segment (approximately 3 minute length in this example), and the lines represent the continuous phase over the segment proportional to the HPT estimated frequency for that segment; different lines are used for adjacent phase estimates.

The overlapping phase functions plotted in this Figure provide a quantitative measure of how coherent the HPT-estimated harmonic is versus time, as indicated by the agreement between adjacent estimates across their overlap region (recall that this example used a shift between segments defined as 50 percent of the segment length, resulting in a 50 percent overlap). Interpreting the phase continuity is not always straightforward. For example, if the amplitude gets small then the phase estimates become less reliable; also, the phase will not be continuous when independent components merge. However, this capability to inspect phases over adjacent segments is offered as a potentially powerful tool for reliably separating signal (coherent phase) from noise (incoherent phase) components. Note also that a similar plot of FFT-estimated phase versus time would not be informative since it is known that biases would be added to the phase as a function of the fractional bin number as the frequency shifted between the integer bin numbers.

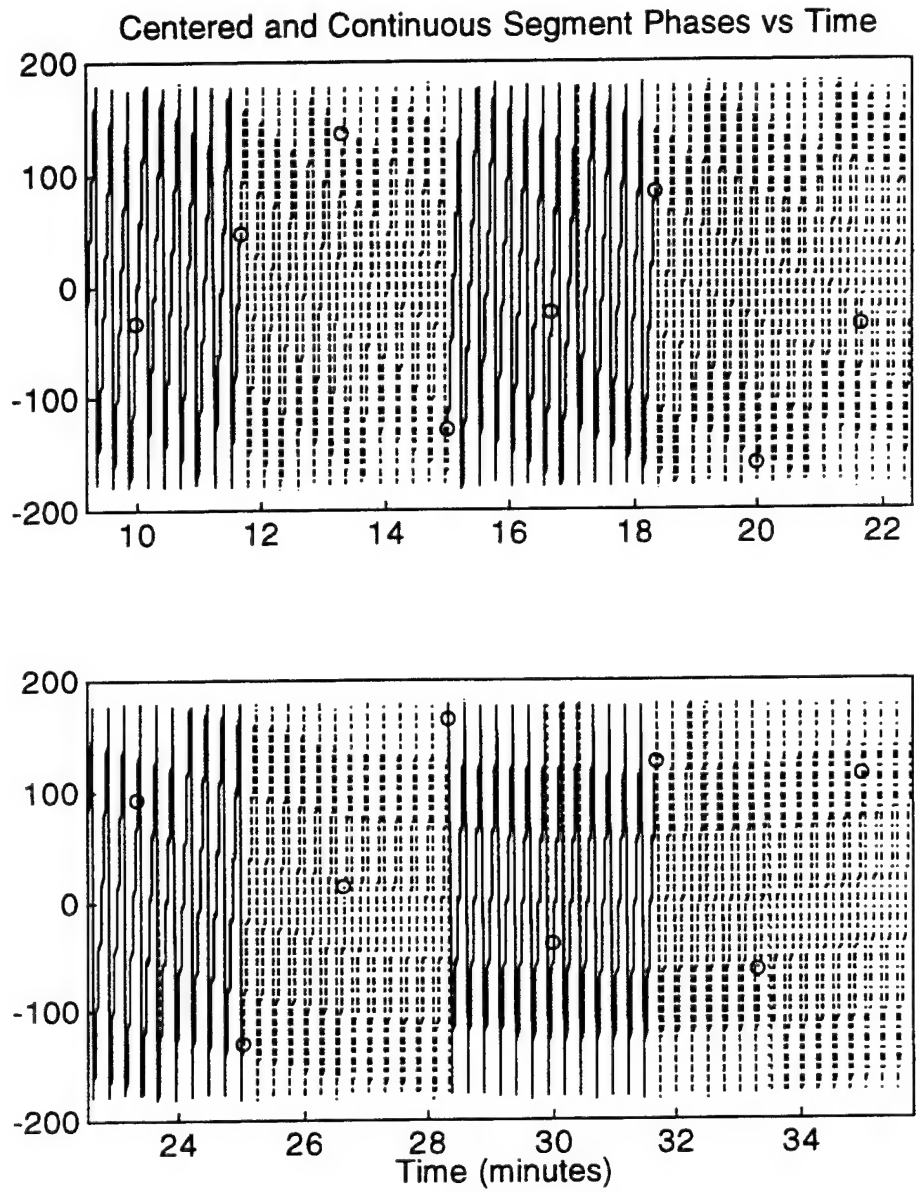

Figure 10b. HPT-estimated Phases versus Time for a Signal with a Linearly-Varying Frequency

The next example HPT analysis of a signal with time-varying parameters is a more difficult signal comprised of two sinusoids with a constant and oscillating frequency that occasionally come very close together. Again, the signal was analyzed using HPT and FFT techniques with the comparable segment lengths and 50 percent shift in start times. Results are shown in Figure 5.11. The HPT results are more informative than the FFT results in two ways: (1) HPT shows two dominant amplitudes, indicating two sinusoids, and (2) both amplitudes are reasonably constant (the filled-in circle at the top of the center figure indicates the two true amplitudes). The instantaneous frequency is shown as the dash-dot line and clearly shows how strongly amplitude variations can affect this function and therefore how unreliable it can be. Figure 5.11 also shows how both techniques have difficulty identifying the two sinusoids when the frequencies come to close together and the segment length (shown in the middle subfigure) becomes only a small fraction of one envelope cycle.

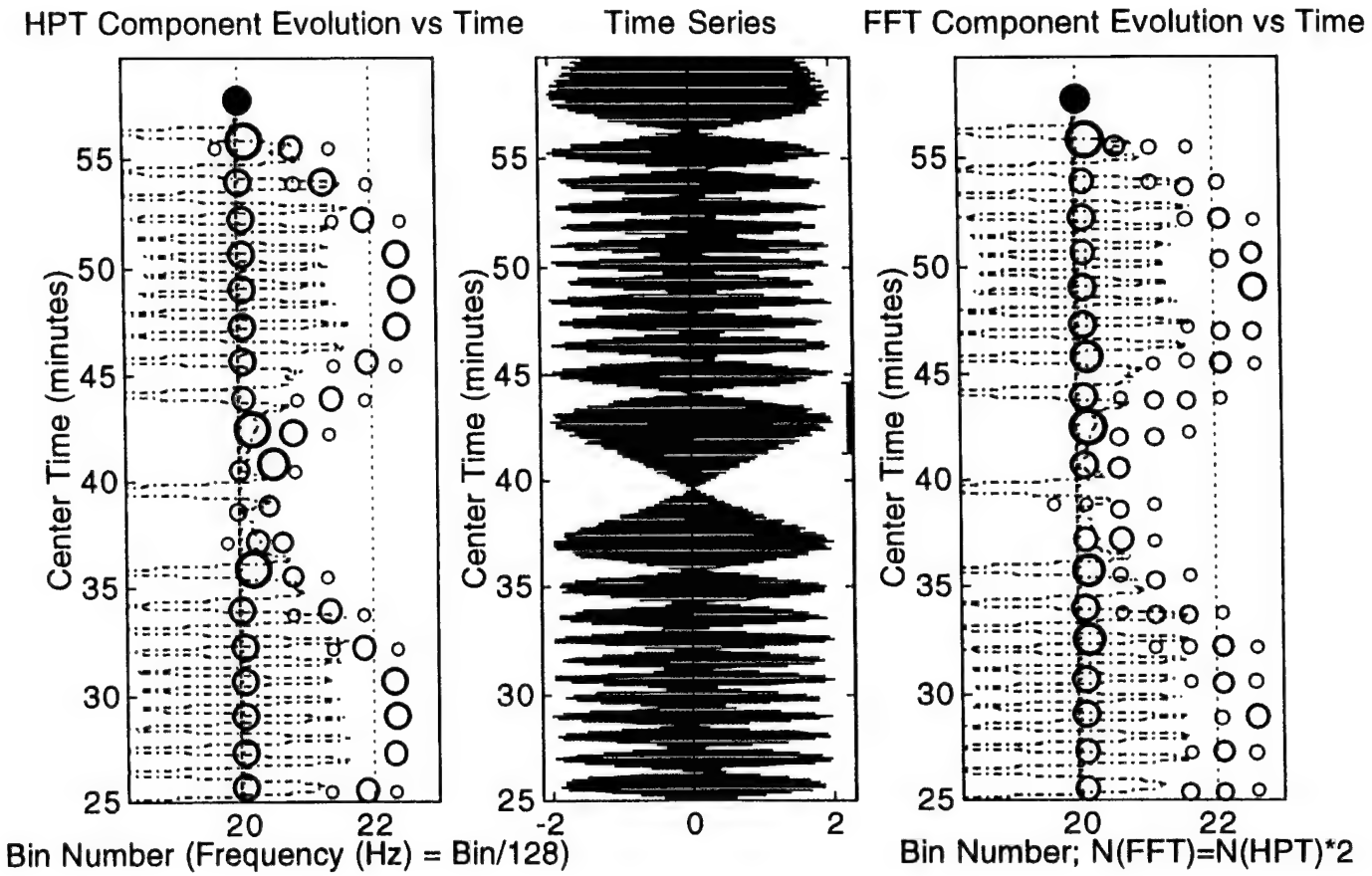

Figure 5.11 HPT- and FFT- Estimated Component Evolution versus Bin Number for Sinusoids with a Constant and an Oscillatory Frequency

Figures 5.12a and b show the HPT estimated bin numbers, amplitudes, and phases for the component sinusoid with the oscillating frequency. The estimated amplitudes are very close to the unit amplitude except for the segment at the node of the envelope of the time domain signal where the two component frequencies are very close. But most importantly, the phases show excellent consistency between adjacent segments, indicating that this nonstationary component was truly a coherent harmonic component and not a convenient fit from the mathematics. It is noted that equivalent plots for the bin number, amplitude and phase of the constant frequency component sinusoid showed excellent agreement with the true values, but are omitted here for brevity.

These examples used a large overlap between successive analyzed segments, whereas standard practice for spectral ensemble averaging uses independent segments (except when windows are used and a 50 percent overlap recovers information "lost" by the tapering at the ends). This large overlap is not required by HPT, so why was it used? A high degree of overlap simply allows for increased continuity when highly nonstationary signals are under study, particularly the frequency variation and the phase continuity aspects. It certainly could be postulated that such overlap maximizes the continuity due to sharing the same signal information between successive analyses. But consider that phase plots based on successive FFTs with large overlap rarely exhibit such continuity - clearly because leakage biases the estimated phases and

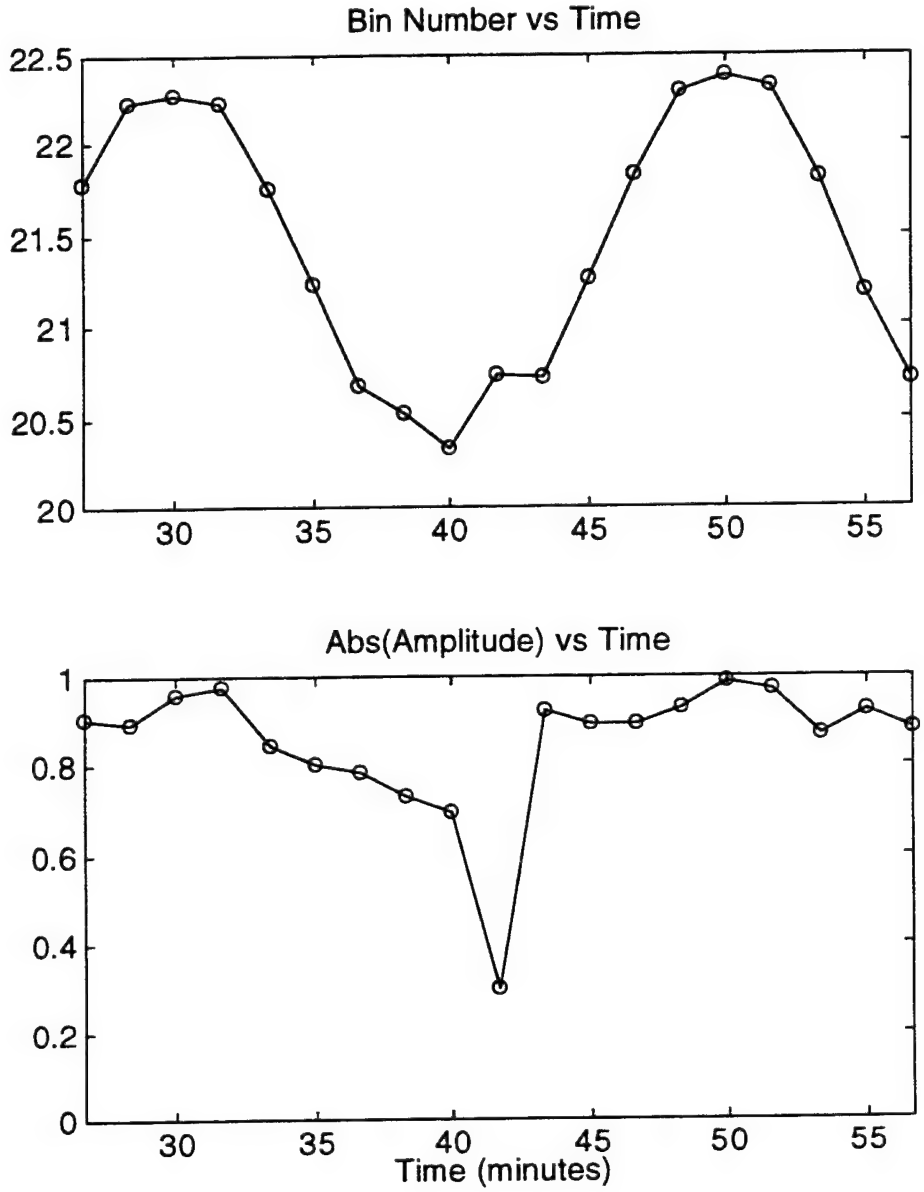

Figure 5.12a. HPT-estimated Bin Numbers and Amplitudes versus Time for a 2-Component Signal with a Constant and an Oscillating Frequency

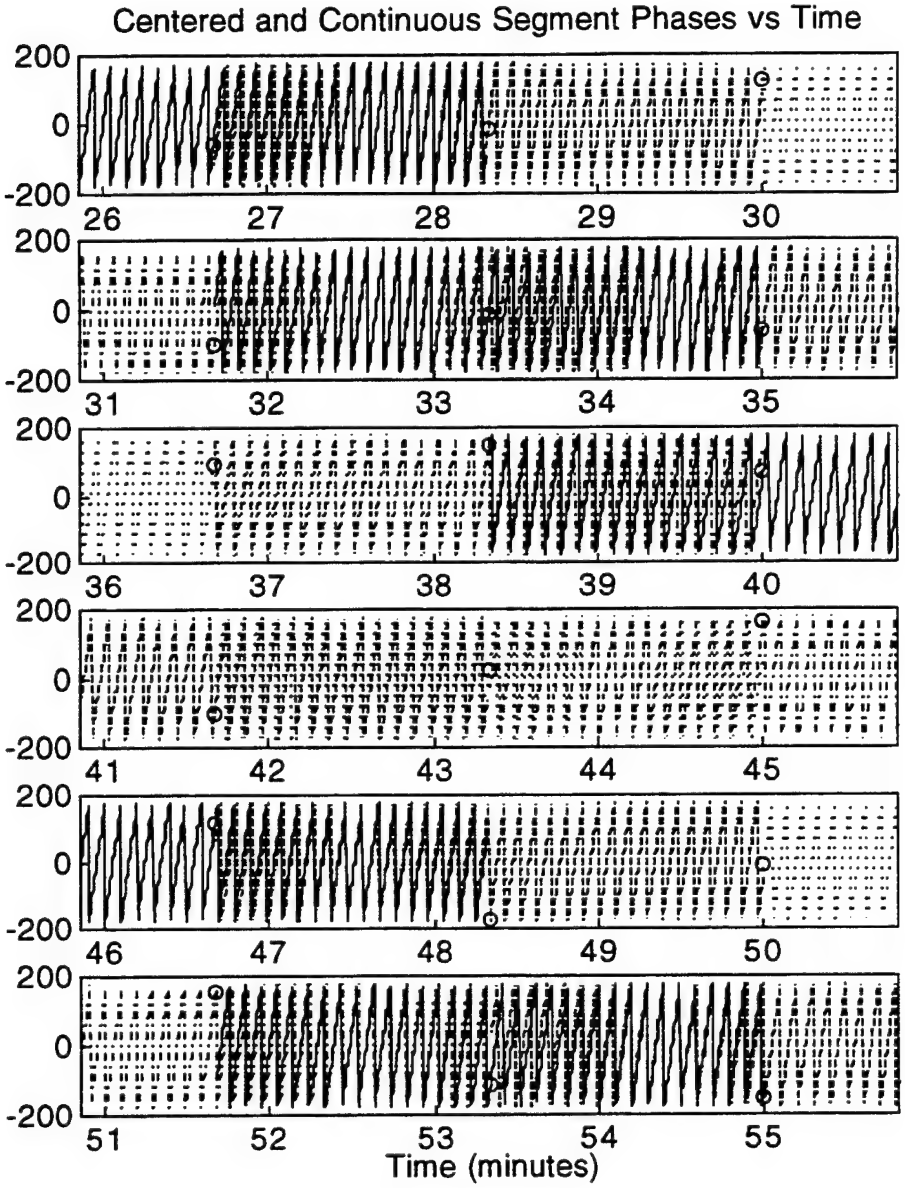

Figure 5.12b. HPT-estimated Phases versus Time for a 2-Component Signal with a Constant and an Oscillating Frequency

the integer frequencies are usually wrong. The practice adopted in most of the HPT numerical studies was to use a large overlap, which made it easier to visually track the evolution of frequencies in general applications (like ocean waves). Regarding the validity of this practice, the next subsection presents one study demonstrating that high overlapping (80 percent) does not bias successive phases; certainly a more formal parametric study could be done along with an analytical study of the reduction in degrees of freedom for each HPT estimate. Before leaving this issue, the reader is reminded that strict independence between estimates is not a universal goal; for example, directionality estimates from an array assumes and requires dependent signals.

In summary, the primary objective of these last two subsections was to demonstrate that HPT can successfully model nonstationary deviations. The objective was not to exhaustively quantify the performance of HPT to a wide range of such signals. This success was very important to establish before attempting to model real-world signals where the signal characteristics are not known and may well include such nonstationarities even over short segments as discussed in Chapter 2. While the discussions in these last two subsections have shown that the technique may not always be a strictly linear operator, they have shown that such variations will not cause the technique to diverge. It was not considered pertinent to the main objective of this ocean wave study to investigate additional types or combinations of analytical nonstationary signals or to add noise.

5.2.5 Representative Analysis of White Noise Signal

It is always necessary to qualify and quantify how any signal processing technique performs in the presence of noise. Band-limited, additive white noise (either with a Gaussian or uniform distribution) is by-far the most prevalent choice in the literature for numerically investigating the behavior of signal processing techniques for "real world" signals, and it is used in that context in this and the next subsection as well. Following this standard practice, the next subsection will examine a stochastic signal defined as the same deterministic multicomponent signal used in the first subsection summed with white noise which has been low-pass filtered to approximately the same bandwidth as the deterministic signal. The resulting data vector is the linear sum of these two vectors.

If the Harmonic Phase Tracking technique was a linear operator, then it would be expected that the estimated parameters for this signal-plus-noise summed vector would be a linear sum of results from independently analyzing the [deterministic] multicomponent and the [stochastic] noise vectors. Unfortunately, although the technique is linear at each iteration (using the total least squares basis matrix), it cannot be unequivocally categorized as a linear operator in some situations. Certainly, the Harmonic Phase Tracking technique satisfies the first criterion for a linear operator $\mathcal{L}$:

$$\mathcal{L}\{ax(t)\} = a\mathcal{L}\{x(t)\} \qquad 5.2$$

The technique does not, however, always satisfy the second criterion for a linear operator:

$$\mathcal{L}\{x_1(t)+x_2(t)\} \neq \mathcal{L}\{x_1(t)\} + \mathcal{L}\{x_2(t)\} \qquad 5.3$$

The reason for this is the time-frequency ambiguity, which must be accounted for since the true signal characteristics are usually not known *a priori*. As Section 5.3 will show, for two closely-spaced sinusoids the results can show either one or two sinusoids, depending on the length of the data vector and the phase of the envelope.

Having concluded that the technique is not strictly linear, it can next be questioned whether it is a linear operator if the time-frequency ambiguity is eliminated. The answer here is that, in these cases when the HPT data vector is sufficiently long (for high frequency resolution), the technique is linear. The numerical results in the first subsection confirm this statement and both of these linearity criteria.

But because it is not possible to always know the minimum frequency separation for the signal or noise frequencies, it must be concluded that the HPT-estimated parameters for a vector defined as a signal plus noise is not the linear sum of independently analyzing the [deterministic] multicomponent and the [stochastic] noise vectors. The objective of this

subsection is to investigate how the HPT technique handles bandlimited white noise alone, so that the information can be used in the next subsection when the total vector is analyzed.

A Gaussian-distributed white noise vector was created using the intrinsic randn function in MATLAB 4. This was low- and high-pass (using a cutoff frequency of half the Nyquist frequency) filtered. In the first noise study, four sequential but independent (no overlap) segments of 200 points were analyzed. Representative bin domain HPT results for two bin regions are shown in Figure 5.13. The four independent analyses are identified by four different line types and the letters a through d next to the peak of each component amplitude.

The first observation is that the *frequencies* are essentially randomly located, with little consistency between the independent analyses. This is expected, since there are no "true" sinusoids in a white noise signal for the technique to find. [Note that HPT does its best to fit the signal and does converge to a "best" set of frequencies for each segment.] Ensemble averaging, or simple inspection, of *frequencies* would conclude that there were no deterministic (i.e., repeatable) component frequencies among these segments (except for an occasional apparent consistency at low frequencies that would be removed if additional segments were analyzed). Also, note that *amplitudes* among similar bin numbers are inconsistent among the four cases. In contrast, if sufficiently-long independent

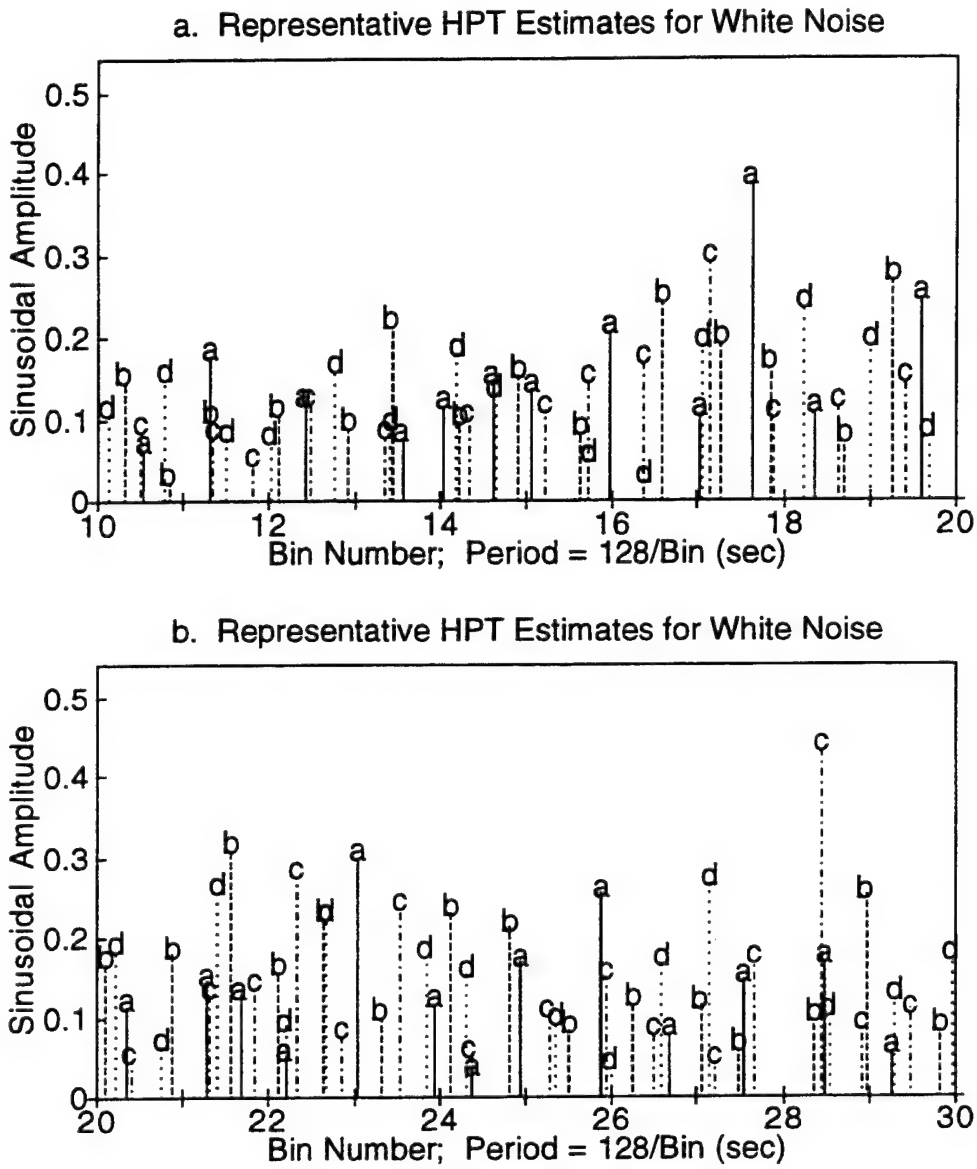

Figure 5.13 HPT Estimated Components versus Bin Number for Independent Segments of a White Noise Signal

segments of the deterministic example signal in the first subsection were analyzed and compared, the frequencies and amplitudes would be essentially identical (see the next subsection). This is another key property of Harmonic Phase Tracking for stochastic signals: *ensemble averaging allows for reliable separation of the deterministic components from the noise.* The inconsistency among the estimates shown in Figures 5.13 is independent of segment length, which in itself provides clues that the underlying signal does not have a deterministic component at any time scale.

A second analyses of a white noise signal was conducted to determine if a high percentage of overlap between adjacent segments would incorrectly show phase coherences. As before, a low- and high-pass filtered white noise signal was constructed. Segment lengths of 200 points (approximately 3 minutes) were used, with a shift of only 40 points between successive analyses; in other words, a very high 80 percent overlap was used.

Figure 5.14 shows one representative span of bin numbers from the HPT estimates. This figure is quite unlike the previous figures that displayed sequential bin numbers. There are no obviously deterministic components. However, note that there are numerous "runs" where the bin number is relatively continuous over finite time spans, and the "run length" is significant and/or larger than the 3 minute analysis length.

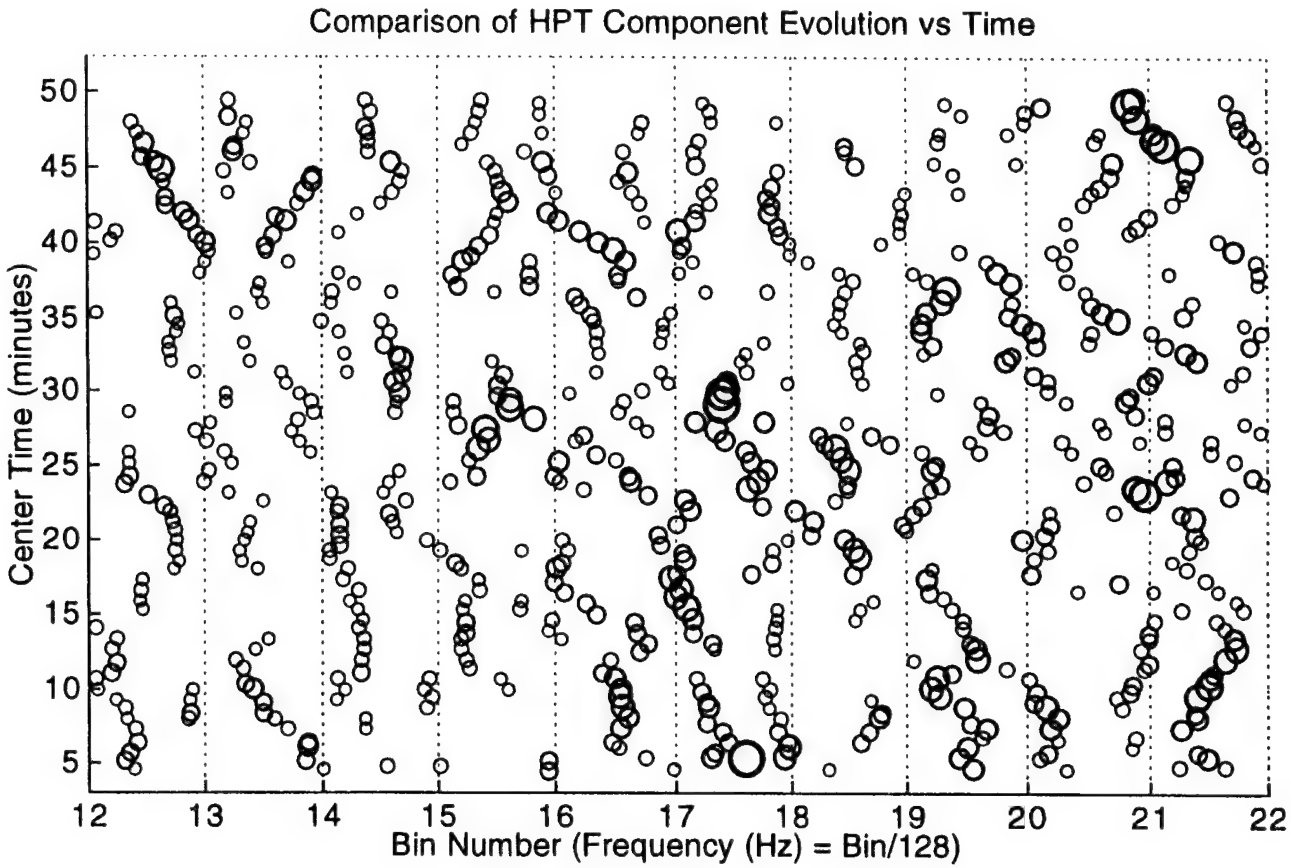

Figure 5.14 Evolution of Bin Number Estimates for White Noise Signal

If this was a signal with unknown characteristics, a fundamental question would be: do these runs represent finite time span coherent harmonics (e.g., a transient sonar signal)? The answer is not clear from inspection of Figure 5.14.

But use of the bin number, amplitude, and in particular the phase evolution plots provides the quantitative information necessary to critically evaluate coherency for any run of interest. For example, consider the run with bin numbers between 15 and 15.5 and times 37 to 45 minutes. At first glance this appears to be a candidate harmonic signal: the run length is considerably longer than the 3 minute segment length, and bin number and amplitude evolutions plotted in Figure 5.15a seem well-behaved and possibly coherent. However, the phase evolution shown in Figure 5.15b shows no consistency between adjacent HPT phase estimates, conclusively proving that this bin number series is not a harmonic. This conclusion is evident even with the high 80 percent overlap. This same inconsistency was present for all of the runs examined in Figure 5.14. This behavior is quite different from the consistency evident from the [nonstationary] deterministic signals analyzed in the previous subsections.

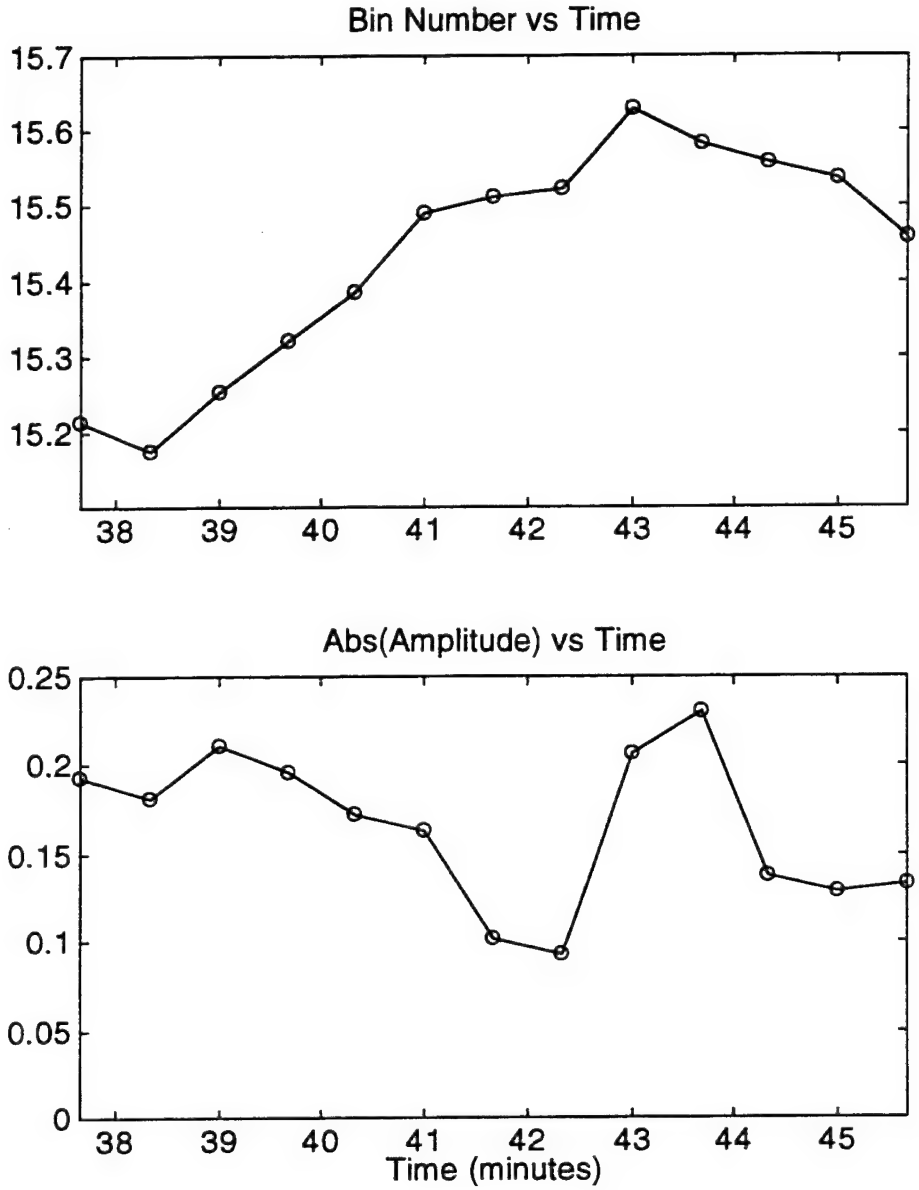

Figure 5.15a HPT-estimated Bin Number and Amplitude Evolution for One Representative Component in a White Noise Signal

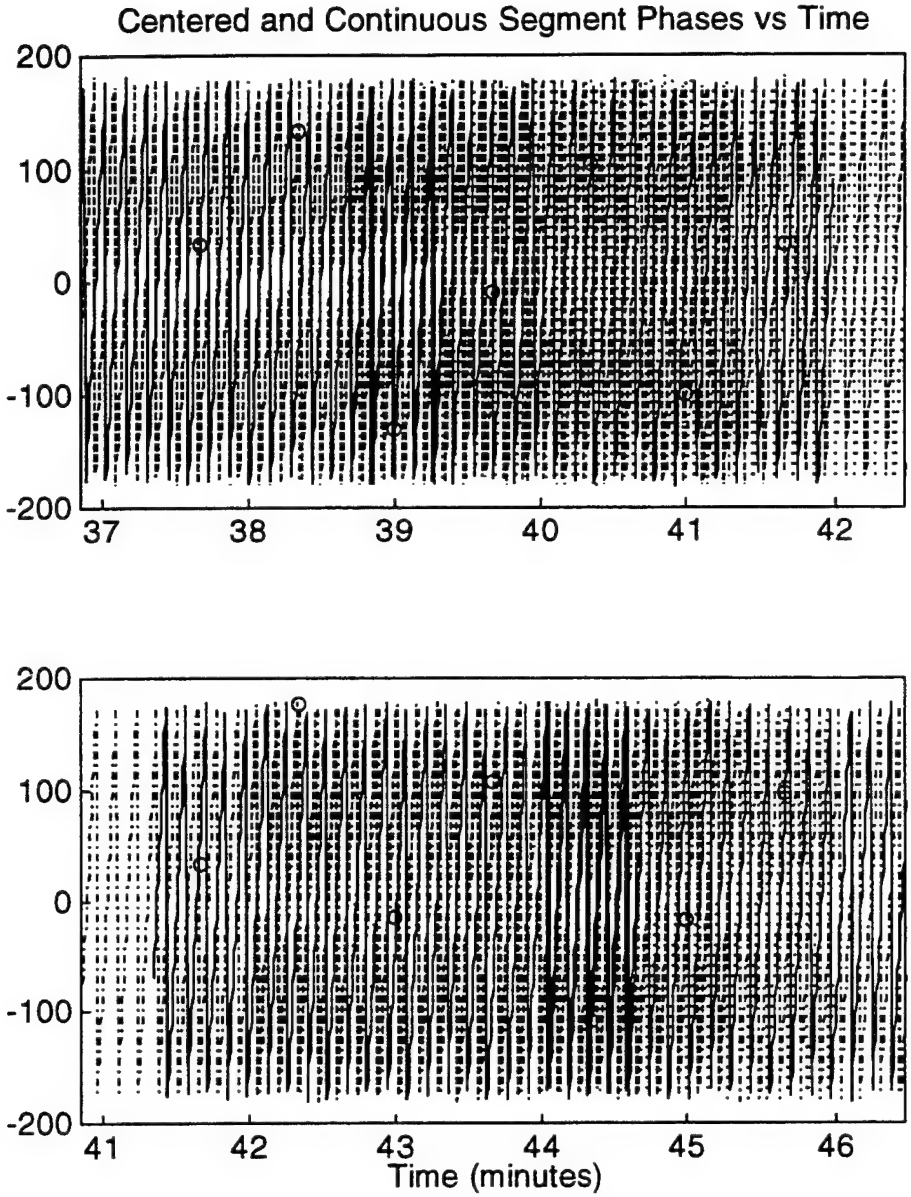

Figure 5.15b HPT-estimated Phase Evolution for One Representative Component in a White Noise Signal

5.2.6 Representative Analysis of Multiharmonic Signal with Additive White Noise

This subsection investigates the behavior of HPT for the standard case of a constant parameter deterministic signal plus noise. A total signal was defined by summing the same deterministic 19-component signal used in Subsection 5.2.2 with a [bandlimited] white noise vector. The root-mean-square amplitudes were approximately 2.5 for the multicomponent signal and 1.5 for the white noise. Five independent segments with the same 360 point length as used in Figures 5.1 through 5.7 were analyzed and are identified as cases a though e.

One representative bin span of all five HPT estimates is shown in Figure 5.16. The following qualitative conclusions can be made regarding the performance of the Harmonic Phase Tracking technique for this combined signal:

- the technique successfully identified only the true deterministic frequencies, as measured by the consistency among the estimates at those [known] frequencies. It is seen that the noise does introduce a small bias in some of the frequency estimates which appears proportional to the correlation (bin spacing) with

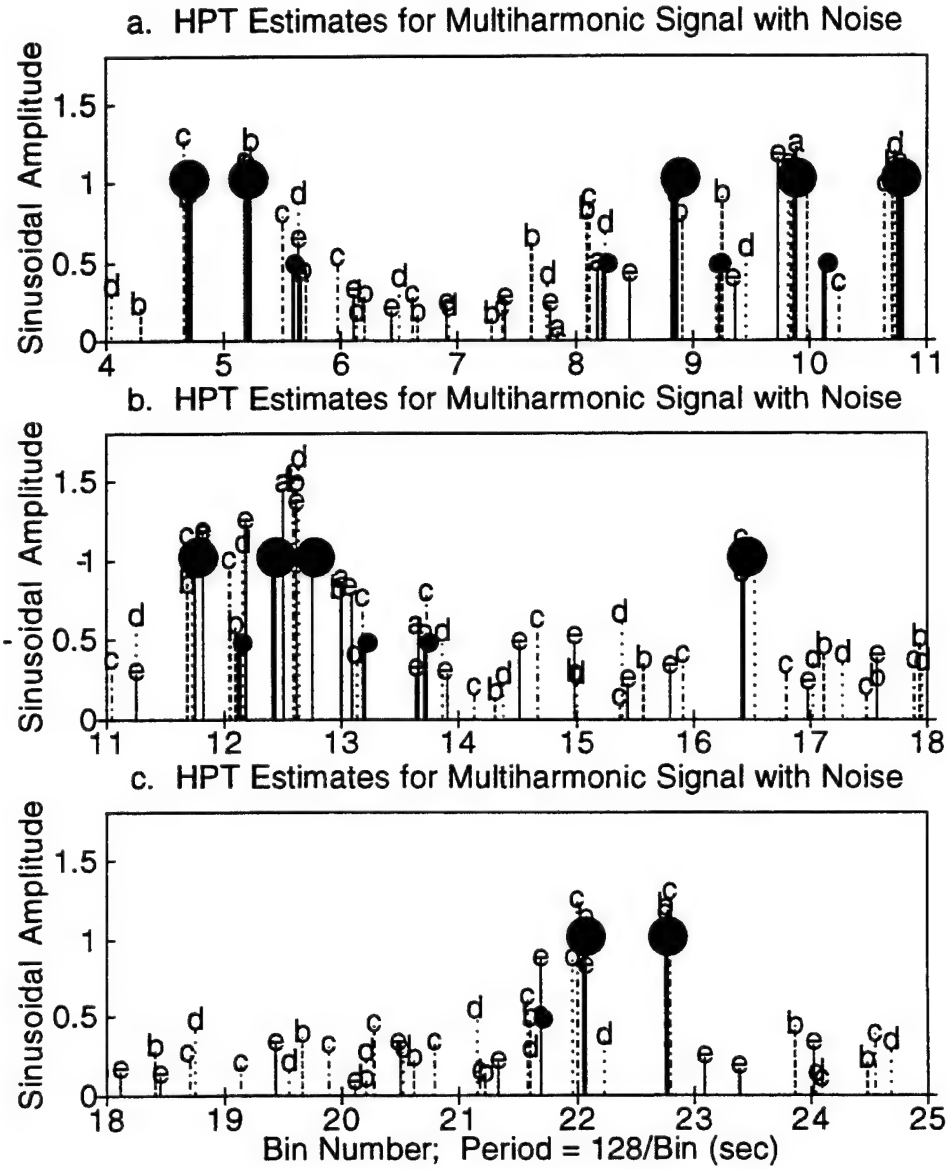

Figure 5.16 HPT Estimated Components versus Bin Number for Independent Segments of Multiharmonic Signal with White Noise

neighboring components; this estimated frequency can be thought of as a "local instantaneous frequency" comprised of a local true component plus a local noise component weighted by their amplitudes squared.

- on the other hand, the technique produced frequency estimates with essentially no consistency at most other frequencies, strongly suggesting that no deterministic components were present there. However, for this choice of segment length and the limited number of samples, there are four bins where there is apparent consistency in frequency, with an appreciable amplitude, without a corresponding deterministic component - at bins 3.5, (perhaps) 7.9, 18.0 and 21.6. Chapter 8 addresses how to best handle and possibly eliminate occasional spurious components like these (one immediate remedy is to increase the number of analyzed segments, but eventually stationarity issues become important).

- both of these first two conclusions show that Harmonic Phase Tracking reasonably estimated the rank of the deterministic part of the signal, i.e., it did not miss any true components and may only incorrectly identify a few noise components.

- while five cases was not sufficient to draw firm conclusions, it is evident that an expected value at each of the deterministic component frequencies would yield a reasonable estimate for the component amplitudes.

In practice, more analyses would be done, and the evolutionary phase figures would be used to identify the coherent, or assumed deterministic, frequencies in the signal. From that best-fit vector, a total least squares analysis will yield the component amplitudes and phases for any segment of interest. This follow-on analysis then allows for any number of studies regarding the signal. Some of these applications are briefly listed below but will be addressed in more detail in later chapters:

- identification of the rank of the deterministic signal
- decomposition of the given signal into a deterministic and stochastic part, which can then be inspected in the time and frequency domains for their behavior (for example, is the noise white or is it concentrated at particular frequency bands?).
- if multiple signals are available from an array, then the relative phases at particular frequencies can be used to estimate incident directions for those components (this is addressed in Chapter 7).
- component frequencies need not be constant in time; inspection of a time-sequence of Harmonic Phase Tracking frequency

vectors would also allow for identification of slowly-varying trends such as a shifting of a system's natural frequency if the system properties are varying in time.

- inspection of time-sequences can also detect the arrival of a "new" component, caused for example by the arrival of a reflected wave or a transient sonar return.

- the component frequencies available from Harmonic Phase Tracking also are a natural choice for predicting future data points. When used in conjunction with the ability to separate the deterministic and stochastic components, this application holds great promise. For example, from studies of actual ocean wave data reported in Chapters 7 and 8, good correlations between the measured and fitted signals are routinely available for four and sometimes more cycles outside of the analysis band. This is well beyond the capabilities of any other known technique.

In summary, these example post-processing applications take advantage of trends in frequencies, amplitudes, and phases available from this technique to learn more about the signal and/or to confirm assumptions used in any such analyses. Undoubtedly, many more applications are possible.

These analyses required between 35 to 80 iterations (approximately 10 to 20 minutes) for each of the 360 point segments and the 55 to 60 components typically estimated from the total signal least squares fits.

5.3 Numerical Aspects of Harmonic Phase Tracking

Two representative studies are presented in this section: the effect of segment length (Section 5.3.1), and dependence on the initial frequency vector estimate (Section 5.3.2). These are the two most important studies to demonstrate that HPT estimates are quite robust and stable. Many other numerical studies, such as the affect of changes to the minimum resolvable bin spacing and various criterion thresholds, were necessarily conducted in the course of developing the MATLAB algorithms but are not reported here since they primarily affect issues such as efficiency and resolution. The effect of noise on HPT-estimated parameters was reviewed in Section 5.2.4.

5.3.1 Data Segment Length and Bin Resolution

The principle objective of the analyses in Section 5.2 was to demonstrate that Harmonic Phase Tracking has the capability to model the full range of signal characteristics expected in ocean waves. Generally speaking, HPT was shown to be a valuable signal processing tool that yields signal

information not available from Fourier Transform and spectral techniques.

Since the signal properties were known *a priori,* the segment lengths for the HPT analyses were purposely optimized to minimize numerical complications. This section extends that capability by focusing strictly on the numerical stability and uniqueness of the HPT estimates.

The only real decision to make in using HPT is to select a length of segment to analyze (other decisions like filtering, sampling rate, etc. are necessary but are independent of the analysis technique). In traditional spectral analysis this requires a compromise between maximizing the length for minimum bin resolution and minimizing the length to: (1) reduce the uncertainty by maximizing the number of stochastic averages, and (2) minimize any possible nonstationarity effects. For some signals it may be possible that any length of segment will return biased results. For example, this could be true for ocean waves for two reasons. First, as stated in Chapter 2, it may be impossible to avoid component averaging at low frequencies and nonstationarity at high frequencies because of the large bandwidth of the signal. Second, it is possible for the spectrum to be continuous at all time scales; if it is, then it is necessary to accept some degree of frequency averaging and not expect to resolve discrete, physically-realizable harmonic components at all frequencies (which is exactly the reason why the concept of the spectrum holds so much appeal).

To investigate the effect of segment length versus bin averaging, a three component signal was defined, consisting of two closely-spaced unit amplitude sinusoids and one isolated smaller amplitude sinusoid. All analyses used a reference FFT length of 64, which results in a minimum true bin resolution (i.e., difference in the number of cycles in the HPT segment) of 0.126 between the first two harmonics. Resolution of this bin spacing by HPT requires a segment length of at least 255 points as defined by Equation C.5. Two different segment lengths were used; the first used 350 points to avoid the ambiguity, while the second used 200 to insure that it did occur. In addition, three different local sections of the signal were analyzed; the first was centered over a crest of the envelope, the second was centered between the crest and next node, and the third was centered across the envelope node. These were chosen to agree with the cases previously discussed in Subsection 5.2.3. Note that relative to the beating envelope period of 1024 points, even the longer 350 point section corresponded to approximately 123 degrees or only 34 percent of a cycle. Thus, even this long segment is not particularly long relative to the beating behavior of the total signal.

Results are summarized in Table 5.3, identified as "crest", "slope", and "node" for the three local sections. Note the following characteristics of the longer 350-point analyses (upper half of table):

Segment Length	Envelope Phase	Bin #1	Ampl. #1	Bin #2	Ampl. #2	Bin #3	Ampl. #3
Exact Parameters:		4.312	1.000	4.438	1.000	6.353	0.250
350	crest	4.315	1.025	4.440	0.975	6.353	0.250
350	slope	4.312	1.000	4.438	1.000	6.353	0.250
350	node	4.310	0.969	4.440	0.969	6.353	0.250
200	crest	4.375	1.827			6.351	0.235
200	slope	4.312	1.000	4.437	1.000	6.352	0.249
200	node	4.301	0.884	4.449	0.884	6.353	0.251

Table 5.3 HPT Analyses of 3-Component Deterministic Signal versus Segment Length and Envelope Phase

- the estimates are remarkably accurate in identifying the signal rank and parameters (usually 2 and often 3 significant digits) for all three local sections. The frequency estimates are particularly accurate.
- The "node" analysis does show a very small 3 percent error in the amplitude estimates. The explanation for this was discussed in Section 5.2.5 regarding the total least squares estimator. Note that this "node" segment only includes ±61 degrees of the envelope, so for this case HPT does not have access to the amplitude information contained at the crest. 3 percent is still considered a generally acceptable error.
- Note that the two beating amplitudes are always identically equal, which is consistent with the characteristics of this envelope per the discussion in Appendix A.
- These analyses required between 14 and 19 iterations.

The 200-point analyses show errors as expected. Note that this shorter segment length corresponds to a very short ±35 degrees of the beating envelope. Observations from Table 5.3 and Subsection 5.2.3 include:

- HPT does not consider the small amplitude variations near the crest to be significant so instead a one best-fit sinusoid was fitted. Note that the frequency is correctly estimated as the mean

frequency for these two equal amplitude true sinusoids ($\bar{f}$ in Equation A2.a). The estimated amplitude for the third sinusoid shows a small bias.

- the "slope" analysis returns excellent estimates of all parameters. [Note: the presently-implemented HPT algorithms do allow for local decreases in the minimum bin resolution if the iteration is well-behaved, which explains how the two closely-spaced sinusoids were accurately resolved even though the segment length was apparently too short.]

- as with the longer estimate, the "node" analysis does show an error in the amplitudes, in this case 12 percent. But they are again correctly equal. And be reminded that this is a very short segment of the time series, so this error is still quite reasonable.

- As with the longer segment results, the frequencies are reasonably accurate for all cases.

The tabular summary presented in Table 5.3 has been informative. But further insight can be obtained if the modified waterfall (i.e., evolutionary) figures from the last section are utilized here. Three examples are discussed.

The first example uses two constant parameter sinusoids at bins 8.765 and 9.00. This requires a minimum segment length of 326 points for HPT to resolve. Instead, 200 point segments were used to investigate the behavior of the HPT estimates when the signal is "quasi-continuous"; a FFT length of 256 points was selected as comparable. The modified waterfall diagram is shown in Figure 5.17, where a 128 point FFT was retained as the reference FFT length for the graphical displays (thus, a 256 point FFT appears to have a resolution of one half bin).

Start by examining the FFT component evolution in the right subfigure. These FFT frequencies are, of course, constant. The predominant amplitude varies in phase with the amplitude of the signal envelope (middle subfigure), but the estimate is poor in the vicinity of the envelope nodes where the phase discontinuity discussed in Appendix A disrupts the instantaneous frequency. While it is conceivable that an experienced analyst could conclude that the beating FFT ordinates at bin 9 and the smaller ordinates at bin 8.5 indicate two underlying sinusoids (indicated by the dashed lines and the fill-in circles at the top), there is little more that can be deduced from this information.

Contrast that to inspection of the HPT information displayed in the left subfigure. As with the FFT ordinates, the HPT amplitude estimates show a beating behavior that follows the signal envelope. But note that the raw HPT frequency estimates very closely track the instantaneous frequency

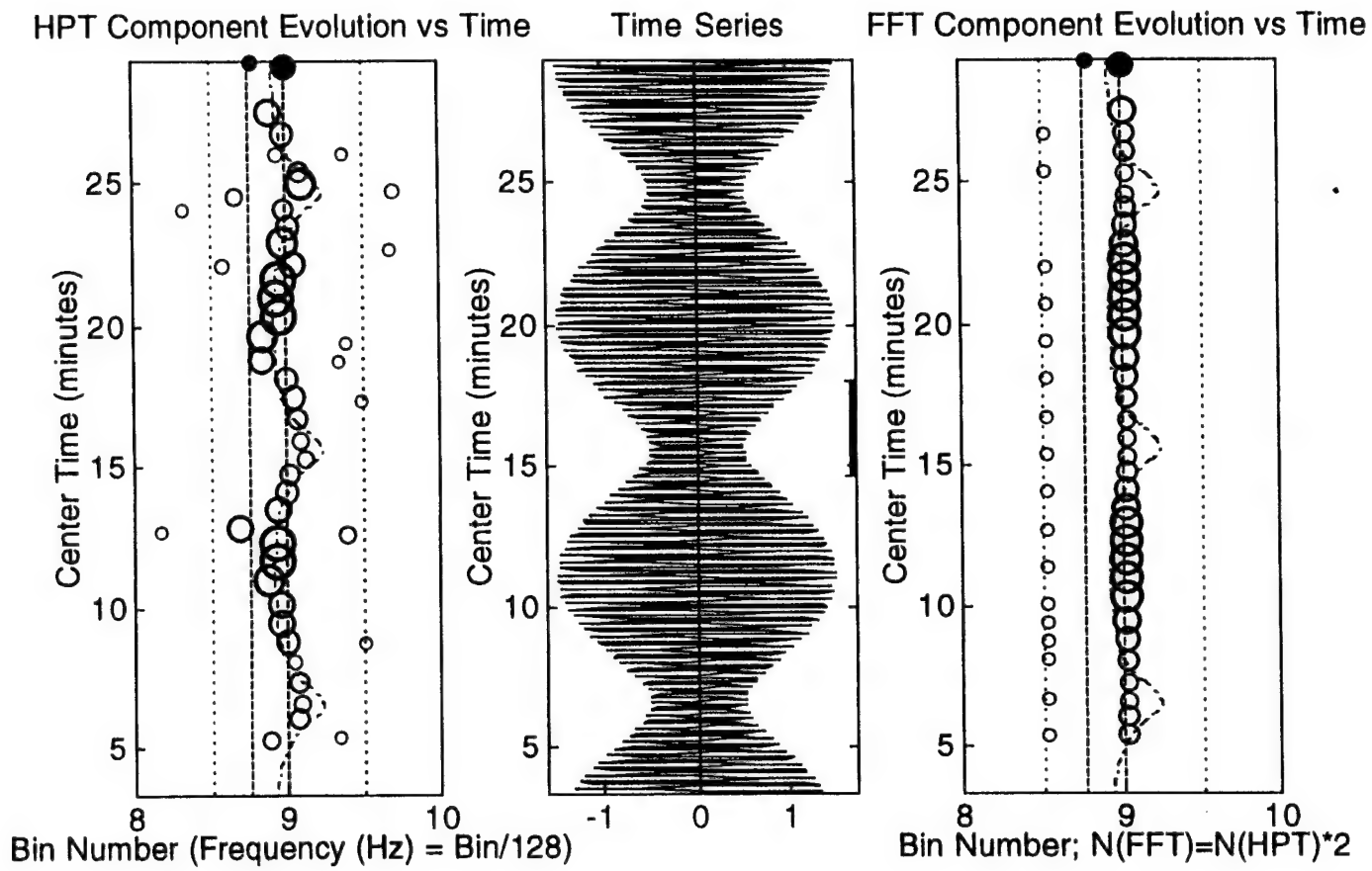

Figure 5.17 HPT- and FFT- Estimated Component Evolution versus Bin Number for Two Sinusoids with Closely-Spaced Frequencies

indicated by the dashed line. This new HPT frequency information is, of course, not definitive in terms of identifying the two underlying sinusoidal components. However, it does recover the instantaneous frequency, and that provides useful additional information about the signal not available from a FFT analysis.

The next example signal uses three instead of two closely-spaced harmonics at bin numbers 14.765, 15.000, and 15.235 (same separation as the previous example). The time series, and HPT and FFT estimates, are shown in Figure 5.18.

The FFT estimates show large amplitudes only where the instantaneous frequency is relatively constant. It would be very difficult to conclude that there were three sinusoids within this bin region.

The HPT estimates are more consistent in amplitude for all segments. The HPT best-estimated frequency very accurately tracks the instantaneous frequency, even through the [partial] phase discontinuities. While it may not provide enough definitive information to identify and recover the three harmonics, the set of HPT estimates is clearly stable, with a behavior that can be physically correlated with the instantaneous frequency.

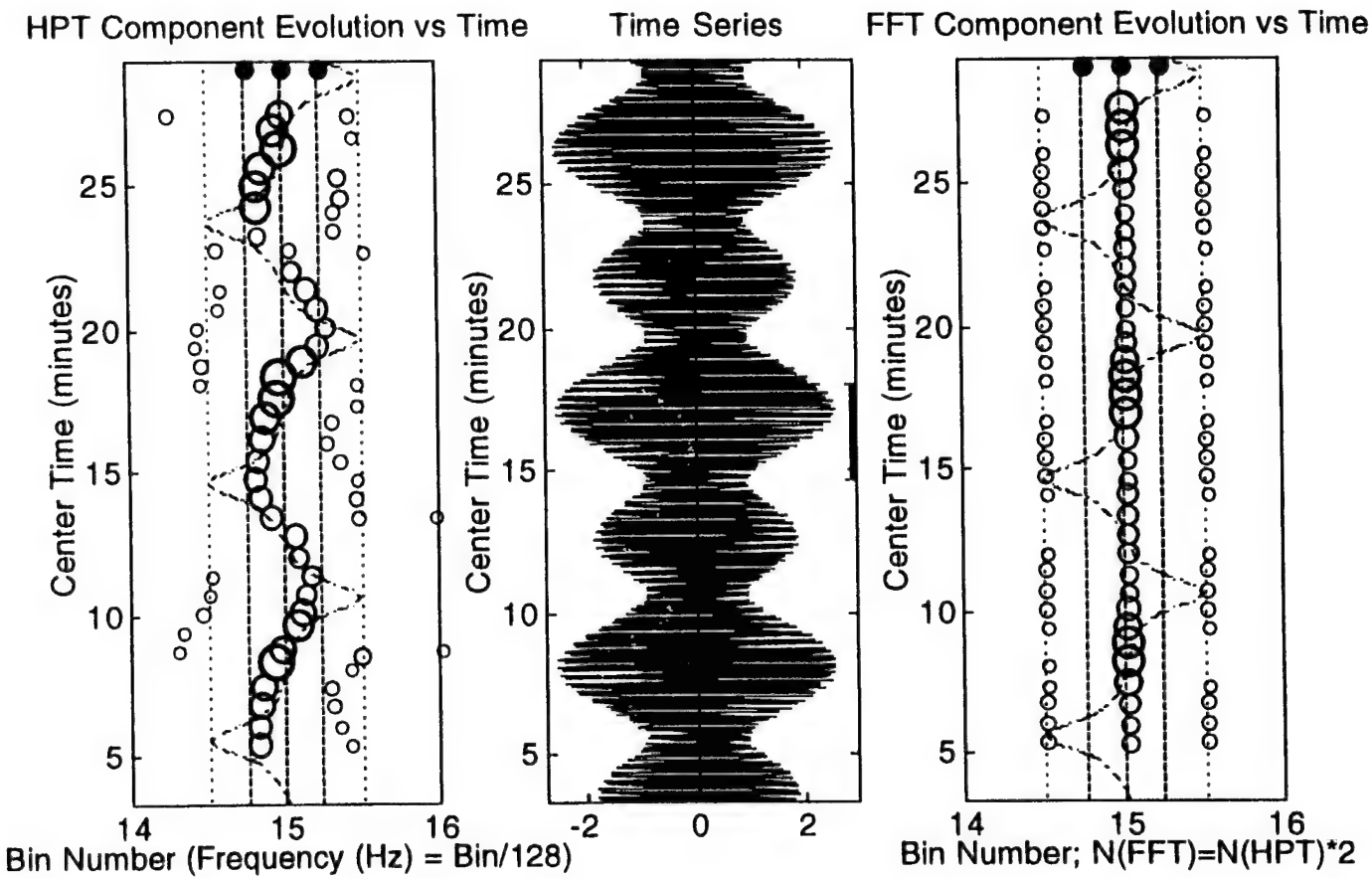

Figure 5.18 HPT- and FFT- Estimated Component Evolution versus Bin Number for Three Sinusoids with Closely-Spaced Frequencies

One final example of three closely-spaced harmonics is presented in Figure 5.19 (same bin spacing and segment length as previous two examples). The motivation for presenting this example can be found from inspection of the HPT estimates in the left subfigure. Examine first the instantaneous frequency (IF) and compare it to the signal envelope in the center subfigure. The time intervals for the IF discontinuities are not equal; one interval is roughly 50 percent longer than the other because of the envelope shape. For this signal the estimated HPT frequency consistently has difficulty converging within the very nonstationary shorter interval. As a result, the sequence of frequency appears coherent only over the longer interval where it continually decreases (downshifts) over time, disappears or at least becomes unreliable, then reappears at a higher frequency 5 minutes later to begin the cycle again. A very long time series would be required to observe that this apparently nonstationary downshifting was actually one part of a longer cycle representing the interactions of three stationary harmonics.

Recall that the IF corresponding to two stationary sinusoids is relatively constant, except for very abrupt delta functions at the envelope nodes. With three harmonics as in this example, the IF as shown in Figure 5.19 is a complicated function which is in effect the sum of three interacting pairs of constant IFs with delta functions. This "sawtooth" behavior of repeated apparent downshifts is interesting because it impacts the ocean wave interpretations in Chapter 7.

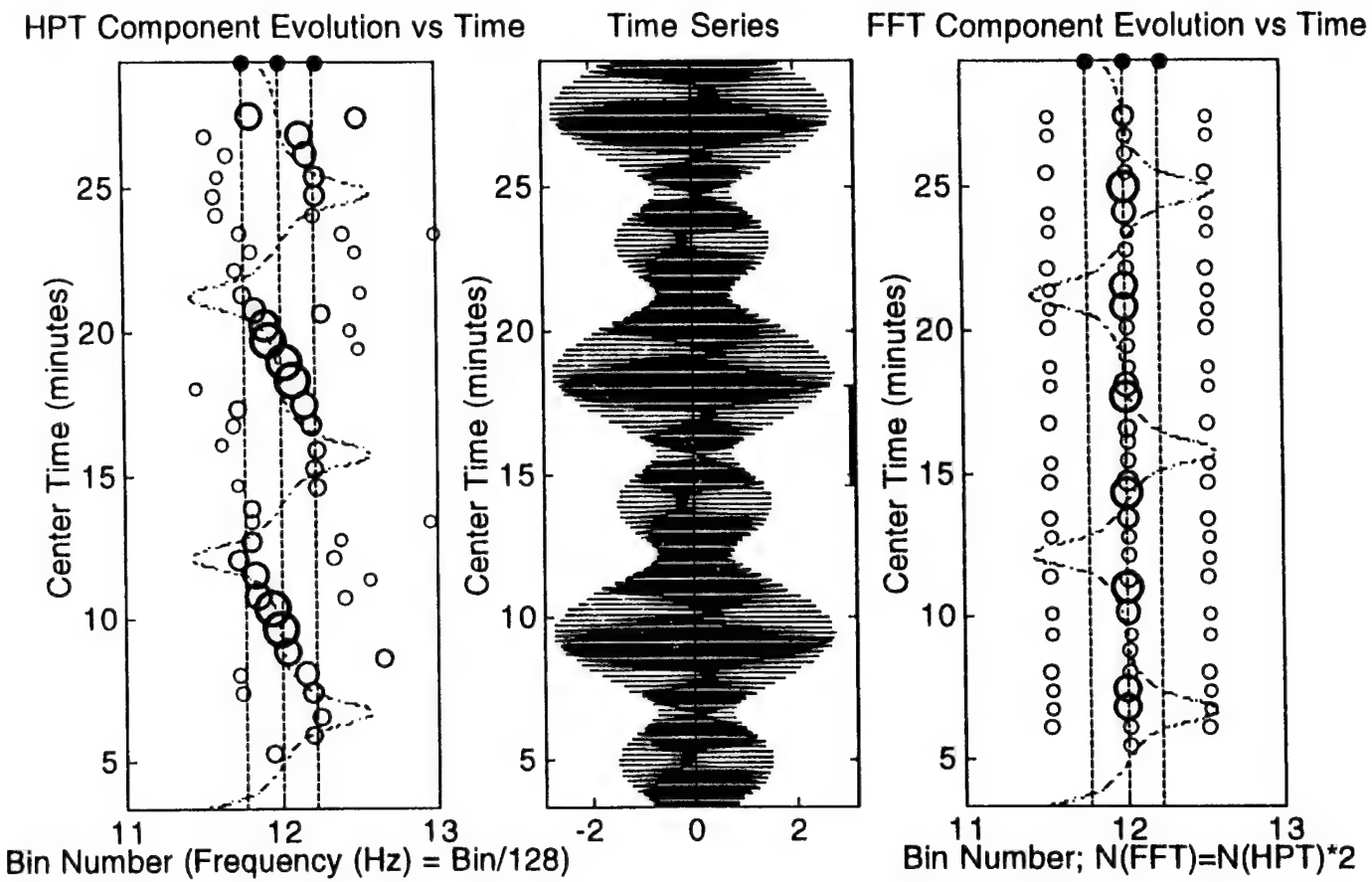

Figure 5.19 HPT- and FFT-Estimated Component Evolution versus Bin Number for Three Sinusoids Showing Sawtooth HPT Behavior

What are some of the major conclusions from this subsection?

1. All Harmonic Phase Tracking parameter estimates are accurate when the resolution (proportional to the segment length) is consistent with the minimum bin spacing of discrete harmonic components.

2. HPT estimates are stable for all signals, and in addition are reasonably accurate for all examined signals except for short segments across envelope nodes.

3. Typically, the HPT frequency estimates are reasonably close to the instantaneous frequency and are therefore related to physical signal characteristics.

4. The tendency of HPT to estimate the instantaneous frequency for closely-spaced components is generally useful but can result in patterns that are difficult or perhaps misleading to interpret.

5. HPT estimates can reveal additional information about the signal that is not available from FFT analyses, such as the rank.

5.3.2 Investigation of Dependence of HPT Estimates on Initial Frequency Vector

The next study regarding the numerical robustness and stability of the technique varied the rank and values of the initial frequency vector to determine what affect that may have on the final best-fit solution. The rank-19 multicomponent of Subsection 5.2.2 summarized in Table 5.1 was used as the basic signal. The same segment length was used. For reference, the previous HPT estimate from that Subsection is summarized in Table 5.2.

The initial frequency vector for the iterations in that reference solution was found from the real transform matrix **R** as discussed in Chapter 4. For this study, that initial vector was modified as follows:

1. The estimated rank was reduced from 19 to 15 by arbitrarily eliminating components at bins 9.23, 10.12, 11.75, and 12.21 (rows 6, 8, 10, and 12 in both Tables).
2. The remaining bin numbers were modified by adding a vector of zero-mean random numbers bounded between -0.25 and 0.25; the rms value of this random vector was 0.13.

The solution using this reduced and modified initial frequency vector is summarized in Table 5.4.

Comp-onent	Frequency		Amplitude		Phase (Deg)	
	Estimated	True	Estimated	True	Estimated	True
1	4.6865	4.6875	1.003	1.00	-42.81	-43.41
2	5.1900	5.1875	0.998	1.00	130.04	129.56
3	5.5850	5.5882	0.497	0.50	-72.52	-73.29
4	8.2528	8.2500	0.443	0.50	-161.73	-161.02
5	8.8798	8.8750	1.007	1.00	98.34	100.19
6	- *	9.2353	-	0.50	-	0
7	9.9008	9.8750	1.251	1.00	137.99	131.13
8	10.0792*	10.1250	0.761	0.50	-38.83	-52.38
9	10.7661	10.7647	1.060	1.00	43.43	44.50
10	11.7488*	11.7500	1.015	1.00	-121.54	-120.23
11	12.1109	12.1250	0.487	0.50	-34.77	-35.51
12	12.4176*	12.4118	0.985	1.00	-178.32	-179.83
13	12.7534	12.7500	0.995	1.00	-88.84	-89.30
14	13.1828	13.1875	0.495	0.50	63.20	62.05
15	13.7064	13.7059	0.501	0.50	-59.31	-60.39
16	16.4378	16.4375	0.992	1.00	-118.27	-118.65
17	21.6874	21.6875	0.497	0.50	32.56	32.52
18	22.0629	22.0625	1.001	1.00	117.02	117.25
19	22.7652	22.7647	0.997	1.00	-124.23	-124.25

* denote components that were removed from the initial frequency vector prior to the iterations

Table 5.4 Estimated Parameters for Multicomponent Signal Using Modified Initial Frequency Vector

Note first that the final HPT-estimated solution does not include the component at bin 9.23. Evidently, the iteration process was able to bias the amplitude and frequency of neighboring components enough so that the error in this bin region did not become large enough to trigger insertion of new components. For example, the amplitudes at the next two highest bins were increased by 25 and 52 percent, respectively. The technique did, however, insert components at the other two removed bin numbers where the missing amplitudes were larger and the error term did trigger insertions. This leads to the conclusion that this missing component could be detected simply by adjusting the insertion thresholds. Recall that the threshold values used in the development of this technique were purposely chosen as a compromise between accuracy and efficiency; in these cases, the 15 iterations required for the reference analysis was increased to 35 for this more difficult case. If increased accuracy was required for a given situation, these insertion thresholds could be tightened at the cost of increased computational costs. That was not studied here.

Otherwise, the solution found using this modified initial frequency vector is considered good with respect to bin numbers, amplitudes and phases.

Many other studies have been performed which have not been reported here. These experiences demonstrated that HPT dependably arrives at a consistently estimate for virtually all signals.

5.4 Summary of Harmonic Phase Tracking Validation

This Chapter has presented many example studies to demonstrate that Harmonic Phase Tracking is capable of modeling a diverse set of real-world signals. In the process, many strengths and some weaknesses of the technique were discussed. Sometimes these characteristics applied to all signals, while other times they applied to only particular types of signals. This section is presented to organize the conclusions and observations made throughout this Chapter to help minimize any possible confusion. Key findings in each numbered paragraph are underlined.

1. The discussion in the last subsection clearly illustrates the point that although this technique is adaptive and iterative, the converged estimates are inherently asymptotically accurate. Better or worse accuracy is available by modifying the convergence thresholds. Therefore, the primary focus of this Chapter has been on the *qualitative* performance of the technique. Since the selected threshold values are compromises between accuracy and efficiency, the accuracies reported here could be considered as minimum accuracies which could be improved.

2. When applied to discrete, deterministic, constant parameters multiharmonic signals, the Harmonic Phase Tracking technique estimates the correct rank and accurate parameters whenever the

segment length is long enough to provide adequate bin resolution. Thus, it is unbiased in these applications.

3. The technique is insensitive to linear variations in sinusoidal amplitude regardless of the magnitude. This is an interesting complement to how Fourier Series models this type of signal. However, it interprets a single sinusoid with nonlinear amplitude variations as a pair of beating sinusoids with an envelope period much longer than the segment length. Both of these apparently different interpretations can be traced to the time-frequency ambiguity issue, so it can be argued that neither are "wrong" when the segment is too short.

4. The technique can successfully model a sinusoidal component with a varying frequency. Generally, the best-fit frequency estimate to several closely-spaced harmonics follows the instantaneous frequency. While this has value because it is related to the signal itself, it can introduce interpretation problems.

5. Even when the segment is too short such that bin averaging of components occurs, the technique models multiharmonic signals reasonably accurately. However, the averaging characteristics vary with the phase of the envelope of the components within any given bin. If a time-sequence of segments is analyzed, inspection of the varying amplitude and rank can provide useful information that the segment should be lengthened and the analysis repeated.

6. HPT is capable of robustly modeling closely-spaced discrete harmonics with a finite set of discrete harmonics. Since such closely-spaced components approximate a signal with a continuous spectrum, this establishes that HPT is applicable to ocean waves if this applies.

7. The Harmonic Phase Tracking technique is equally applicable to stochastic signals if ensemble averaging is used. Each realization (i.e., segment) is modeled as a deterministic, multiharmonic signal, and inspection of the results for a time sequence of realizations allows for identification of deterministic (invariant frequency vs. realization) components summed with a stochastic noise component.

8. The phase evolution plot provides a quantitative measure of coherence of any signal component in time. *This is an important new capability not available with a FFT for interpreting whether a suspected component is deterministic.*

9. There are no restrictions on the spectral or distribution properties of the noise component.

10. This is not a real-time algorithm as implemented. Analysis times are typically 15 to 30 minutes for uncompiled MATLAB4 code on a Sun SPARC10 computer.

All of these features of Harmonic Phase Tracking were important to establish prior to analyzing real ocean wave signals because they: (1) are stochastic, (2) have a relatively large number of components with an unknown (possibly infinitesimal) minimum frequency spacing, (3) have slowly-varying amplitude variations, (4) may have frequency variations and variable rank, (5) and may not have white noise characteristics. This Chapter showed that HPT models these signal characteristics and that sequential inspection of the estimates (particularly the phase) provides a useful "goodness-of-fit" measure. These results provide high confidence that Harmonic Phase Tracking will be a robust tool for the analysis of laboratory and real ocean waves in Chapters 6 and 7.

CHAPTER 6

ILLUSTRATION OF HARMONIC PHASE TRACKING USING PHYSICAL SIGNALS WITH KNOWN CHARACTERISTICS

6.1 Chapter Overview

Chapter 5 used analytically-defined signals to numerically demonstrate the robustness and performance of Harmonic Phase Tracking (HPT). This chapter complements those studies by using HPT to analyze one set of full scale tidal data and three sets of laboratory scale mechanically-generated forced waves. In all four cases the characteristics of the signals are (approximately) known *a priori*, so these actual physical signals serve as a valuable "bridge" between the mathematical exactness of the analytical harmonic waves in the last chapter and the uncertainty of the ocean waves analyzed in the next chapter. Non-essential details are omitted. The primary Chapter objectives are to demonstrate: (1) the first performance of HPT for real-world signals, and (2) examples of the type of new engineering information regarding a signal available from HPT.

6.2 Tidal Record Analysis

Before presenting a numerical example, it is instructive to compare Harmonic Phase Tracking with harmonic analysis, which is the technique of choice used to analyze tidal records. The comparison is straightforward. Both use a finite set of constant-parameter harmonics with arbitrary frequencies to model the signal. But harmonic analysis requires *a priori* definition of the rank and the frequencies before finding the best-fit component amplitudes and phases, whereas all of that information is inherently estimated with HPT. Thus, HPT eliminates the potential errors associated with the use of harmonic analysis; viewed this way, perhaps an alternative name for the "Harmonic Phase Tracking" technique could be "total harmonic analysis".

Near shore tides are a very complex phenomenon involving astronomical forces and local bathymetry. While the periods of the major astronomical harmonics can be calculated with great precision, for most sites this information is not sufficient to allow for analytical estimation of tides because of local basin dynamics (Apel, 1987 and Defant, 1961). Instead, local measurements and the astronomical periods are used with harmonic analysis to fit the tidal component amplitudes and phases for a given site, which, depending on the amount of data available, may only be valid for a finite time interval. As a result, the coefficients and phases are generally

not available for most sites, and they could not be identified for this first numerical example. In spite of this limitation, this example still has value because it demonstrates the accuracy of the HPT-estimated periods versus the known astronomical periods.

A tidal sequence for the summer of 1996 at Barlows Landing Beach on Cape Cod Massachusetts was selected. Figure 6.1 shows the 62 day record (500 data points at 3 hour sampling interval) used for the HPT segment length. The results are shown in Table 6.1:

Astronomical Tidal Mode	Symbol	Exact Period (hours)	HPT Period (hours)	Percent Difference
Principal lunar	O_1	25.8190	25.8230	0.0155
Declination luni-solar	K_2	23.9350	23.9547	0.0824
[HPT estimate]		N/A	12.8840	undefined
Elliptical lunar	N_2	12.6580	12.6528	0.0409
Principal lunar	M_2	12.4210	12.4208	0.0017
Principal solar	S_2	12.0000	12.0010	0.0085
[HPT estimate]		N/A	6.2097	undefined

Table 6.1. Comparison of Astronomical and HPT-Estimated Tidal Periods for Figure 6.1

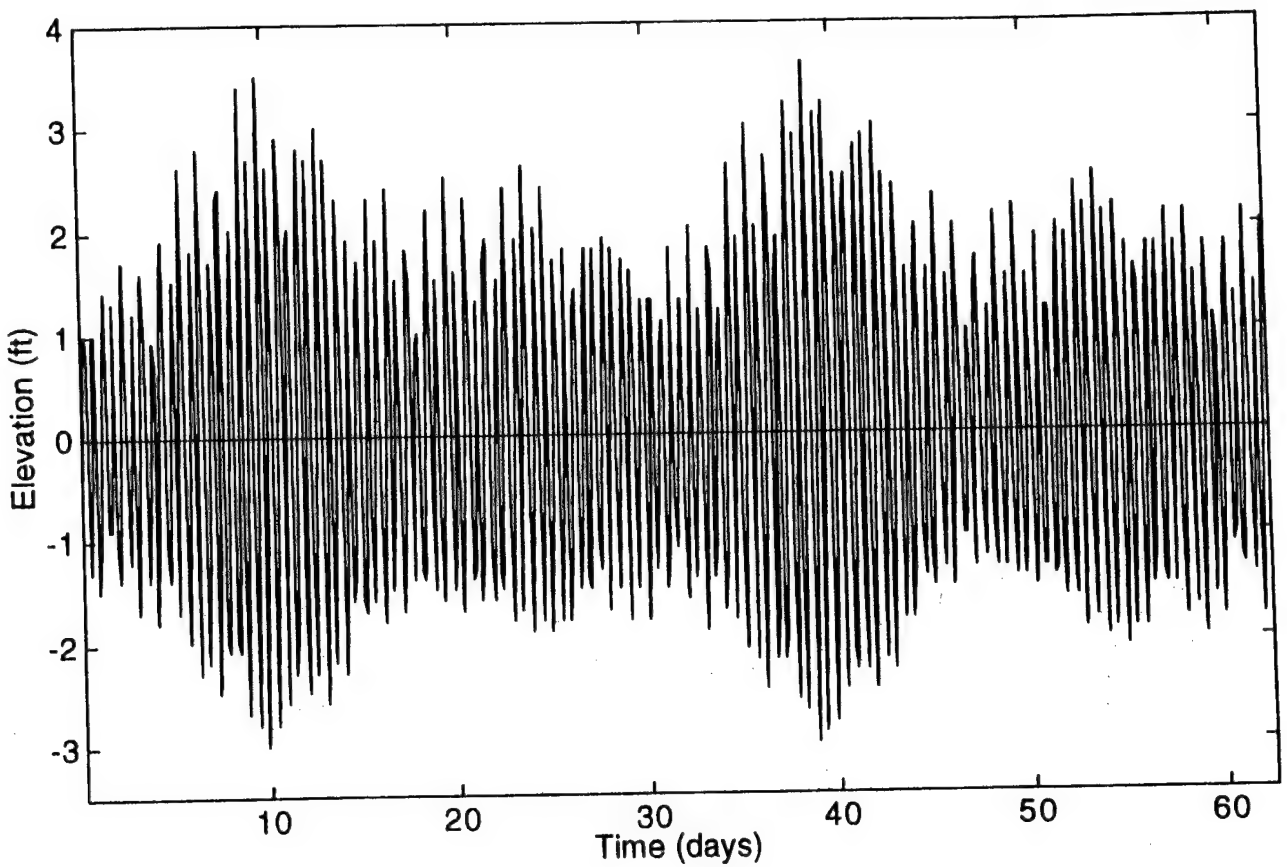

Figure 6.1 Sample tidal record, summer 1996, Barlows Landing Beach Massachusetts

Two conclusions are apparent from inspection of the Table:

1. HPT-estimated periods closely match the known astronomical periods, and

2. HPT detected two additional components. Both of these components have small amplitudes of 0.15 and 0.24 ft, respectively. By way of comparison, the amplitudes for the M_2 Principal lunar amplitude and O_1 Principal lunar components are 2.05 and 0.15 ft, respectively. The source of these additional terms is unknown, but it is interesting that the shortest additional period is exactly the second super harmonic of the M_2 mode, while the longer additional period is essentially equal to the second super harmonic of the O_1 mode (and perhaps exactly equal if a longer data record was used).

Note that the 3- to 4-digit accuracy evident in the HPT periods is not readily available from a FFT analysis; even if an extremely long data segment was available, there are longer astronomical influences that make some component amplitudes nonstationary and therefore introduce bias in the FFT amplitude estimates.

6.3 Rank 4 Laboratory Wave Signal

This is a seemingly simple example. A record was obtained from the U. S. Naval Academy Seakeeping tank of a wave signal made using only four components generated at bins 5, 7, 9, and 11 relative to the 256-point FFT

used to drive the wavemaker, corresponding to periods of 0.45, 0.56, 0.71, and 1.0 seconds. After a suitable start-up interval, a 40 second duration of waves was recorded using a wave staff. The tank dimensions are: 380 ft length, 26 ft width, and 16 ft depth.

The HPT analysis began by high-pass filtering the data and subsequently decimating by 5. The following analysis parameters were used:

number of data points in signal:	205
number of points per HPT analysis:	60
comparable FFT length for figures:	32
number of points per shift:	15
starting data point:	20

This defined a 75 percent overlap, with a total of 8 analyses. The wave signal along with the HPT and FFT results are shown in Figures 6.2a and 6.2b. The format of Figure 6.2a follows Figure 5.9. The decimation by 5 and the change from a 256 point reference FFT used for the wavemaker to a 32 point reference FFT for the HPT analysis results in a new "exact" bin number vector of {5 7 9 11}*(32*5/256) = {3.125 4.375 5.625 6.875}.

Figure 6.2b details the HPT and FFT results. The HPT results clearly show that the signal has a rank of 4, with relatively constant amplitudes. Most importantly, the four HPT bin numbers match the exact bin numbers very closely. The FFT estimates imply two dominant components, but their exact frequencies and any other conclusions are not possible.

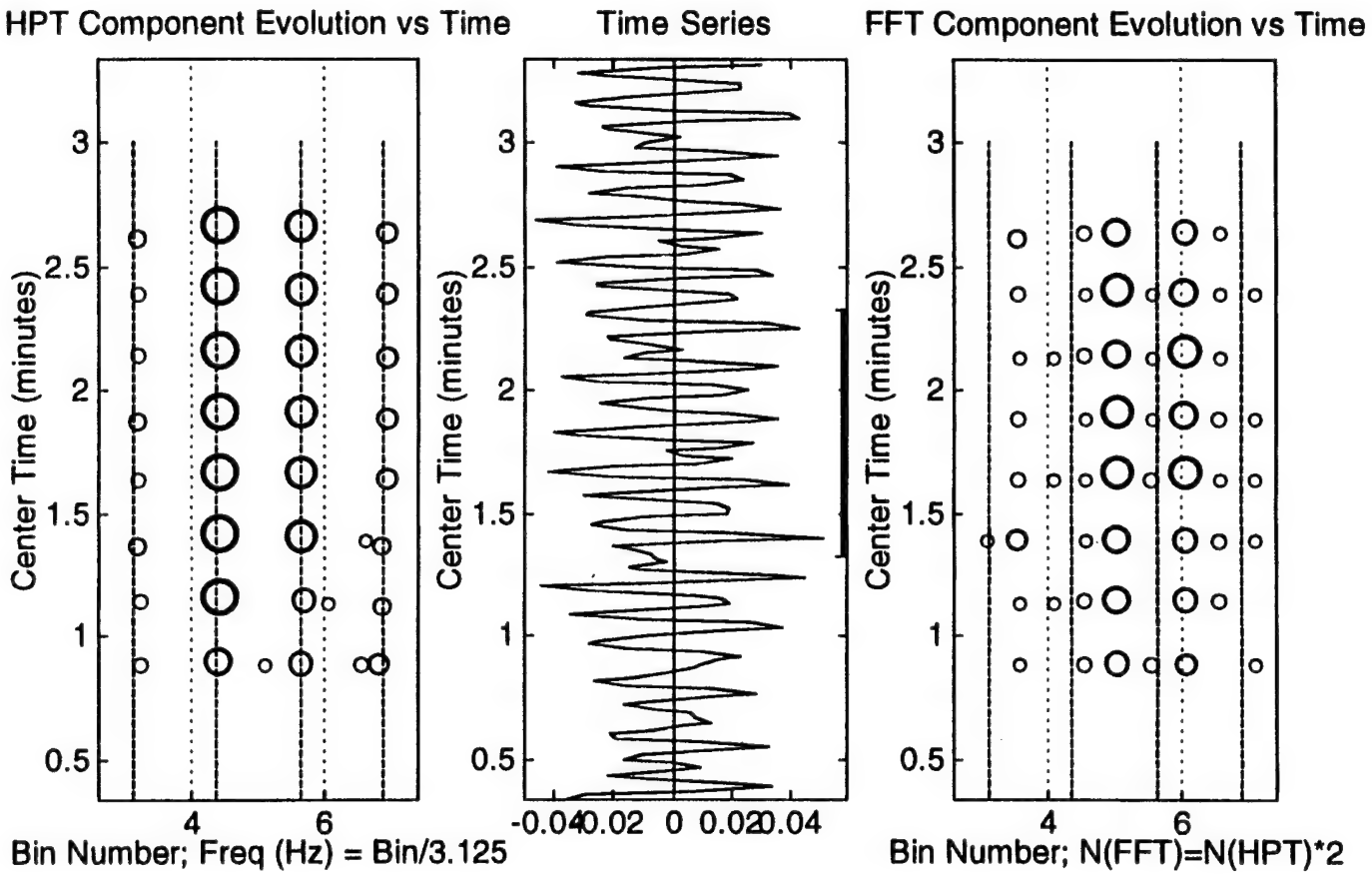

Figure 6.2a. Rank 4 Wave Signal with HPT and FFT Estimates

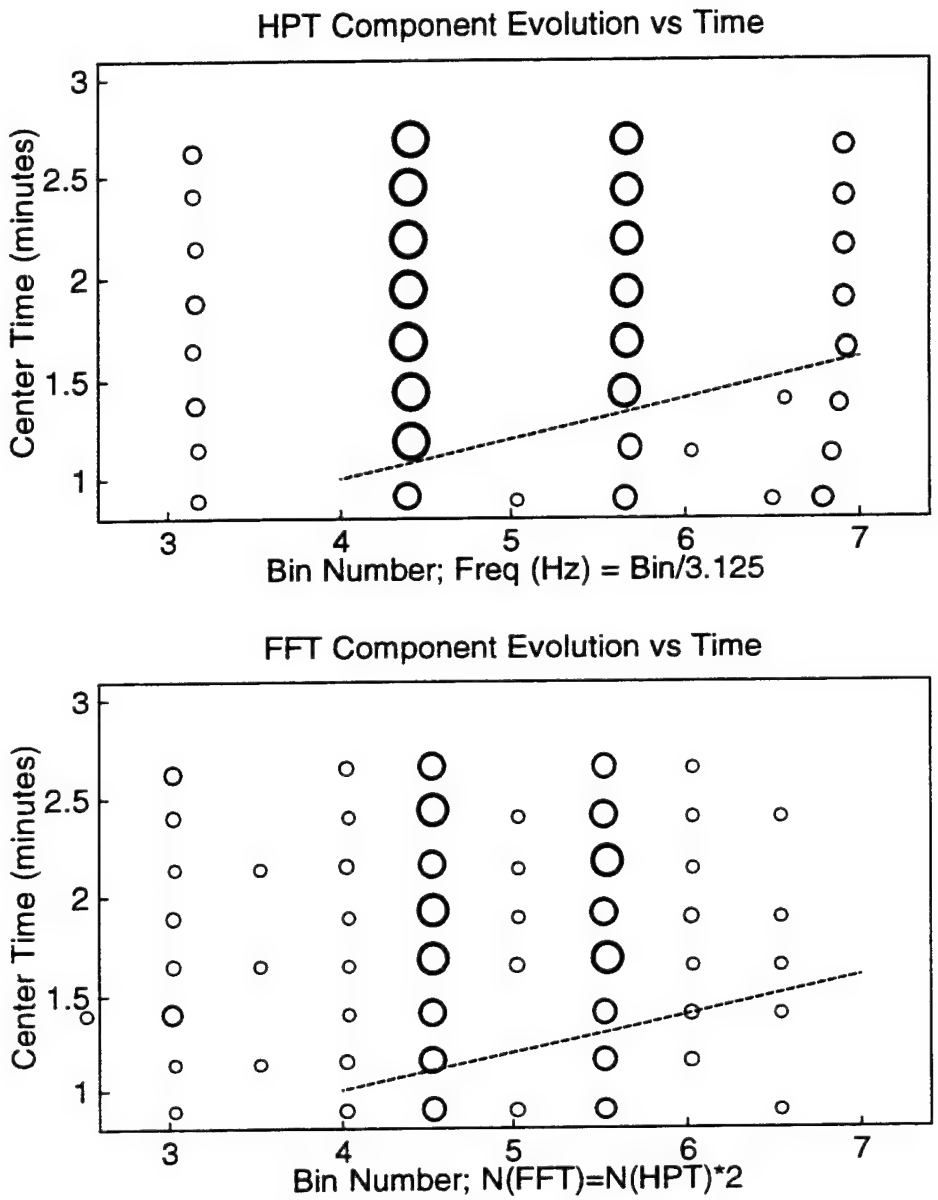

Figure 6.2b. HPT- and FFT-estimated components for Rank 4 Wave Signal

Note that a dashed line has been added to each figure. Inspection of the HPT components in the upper figure shows a clear discontinuity pattern for three of the components that seems to occur at a time that is linearly proportional to the bin number (frequency). The dashed line has been added to approximately mark this demarcation.

This evidence from the HPT analysis does in fact correspond to a physical phenomenon - in this case, energy flux. This is easily shown as follows:

1. The four wave periods (T) are 1.0, 0.71, 0.56, and 0.45 seconds. Since the water depth is 16 ft and the maximum wavelength is $\lambda = 5.12(1)^2 = 5.12$ ft, these correspond to deep water waves.

2. The Deepwater group velocity $C_g\big|_{deep}$ is given by (Dean and Dalrymple, 1984):

$$C_g\big|_{deep} = \frac{\text{wave celerity}}{2} = \frac{\lambda(T)}{2T} = \frac{5.12\,T^2}{2T} = \frac{2.56}{f} = \frac{2.56 * 3.125}{B} = \frac{8}{B} \qquad 6.1a,b$$

   where B is the Bin number.

3. Since any velocity is distance over time, $C_g\big|_{deep}$ is also given by:

$$C_g\big|_{deep} = \frac{\Delta x}{\Delta t} = \frac{100\text{ ft}}{\Delta t} \qquad 6.2$$

   where 100 ft is the distance between the wave gage and the wave generator for these U.S.N.A. tests.

4. The expected slope corresponding to Figure 200 is found by combining Equations 6.1 and 6.2:

$$\text{analytical slope (sec/Bin)} = \frac{\Delta t}{B} = 12.5 \qquad 6.3a$$

5. The empirical slope from Figure 6.2b is found from inspection using the [approximate] (4,1) and (7,1.6) end points of the line:

$$\text{empirical slope (sec/Bin)} = \frac{\Delta t}{\Delta B} = \frac{(1.6-1)(60)}{7-4} = 12 \qquad 6.3b$$

The consistency between Equations 6.3a and 6.3b strongly suggests that this discontinuity measures the fact that the full energy at each frequency arrives at progressively later times due to the dispersive nature of water waves. This phenomenon repeats ad infinitum during any test, making laboratory wave data "piece wise stationary" between the arrival times. This introduces bias into laboratory data that is difficult to eliminate (for multiharmonic waves) and difficult to detect using spectral analysis, even in cases like this where the excitation is purposely defined at Fourier integer bins. In practice, analytical expressions (Equations 6.1 and 6.2) and the experience of the tank operators are used to subjectively define the time interval for the "most stationary" data window. Thus, the individual component wave information from HPT in Figure 6.2b is not presently available and could be useful for laboratory studies.

Continuing with this examination of the data in Figure 6.2b, Figure 6.3 illustrates the HPT parameters for the highest frequency component wave.

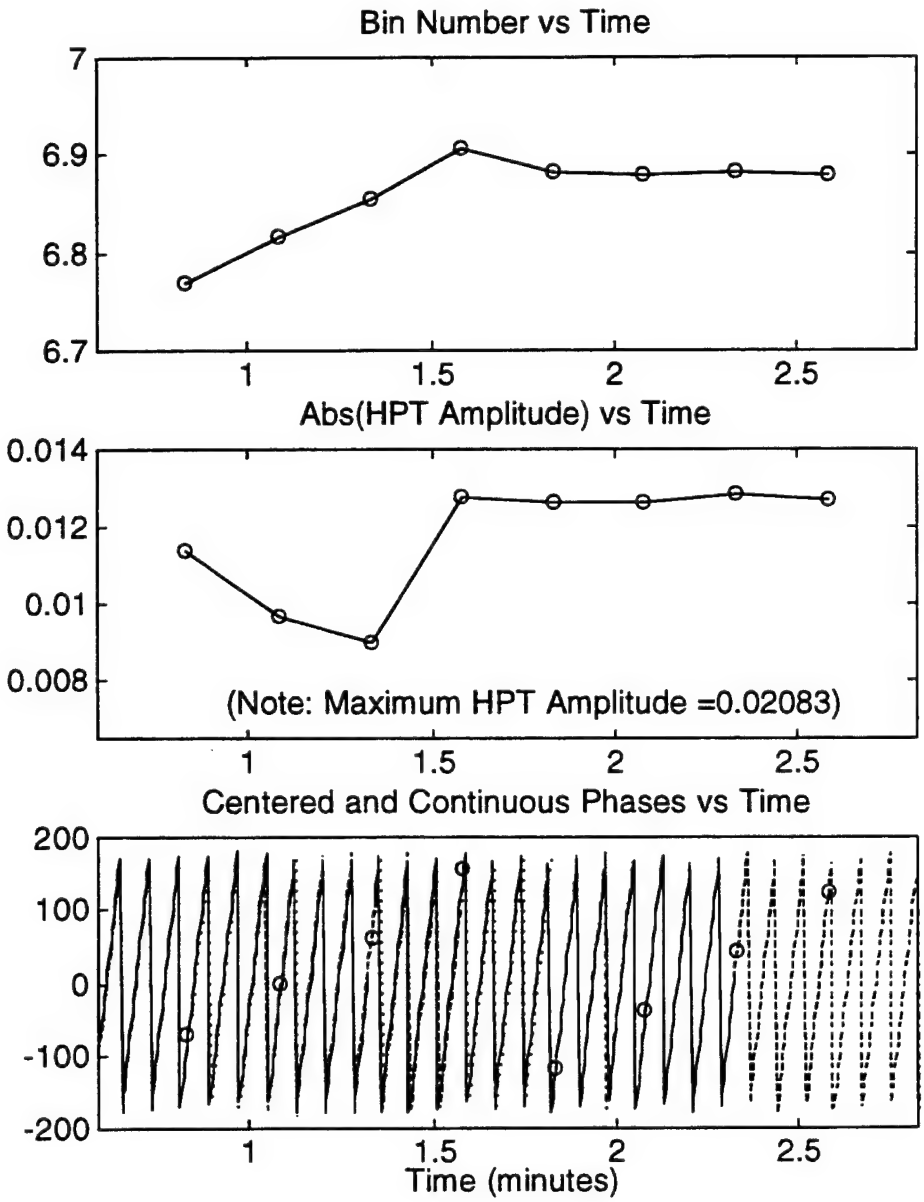

Figure 6.3. Example HPT-estimated Parameters for the Rank 4 Wave Signal

The frequency and amplitude are seen to achieve stationarity after 1.6 minutes, confirming the previous conclusions. The fact that the phase is consistent even during the build-up is not a surprise since the waves are physically present before stationarity is achieved, but just with a lower amplitude.

6.4 Pierson Moskowitz Laboratory Wave Signal

The dispersive energy flux and the recurring reflections of water waves in a closed laboratory basin are not the only complications that experimentalists must consider. For reasons not well understood, the spectrum of the measured waves in many wave basins does not correspond exactly to the FFT-based spectrum programmed into the wave generator (this is true even with the convenient use of orthogonal Fourier frequencies). It is common practice to simply subjectively iterate the program until the measured waves are acceptable (Goda, 1985).

For many laboratory tests the Pierson Moskowitz spectral form is used to represent fully developed, unidirectional, wind-generated ocean waves. The equation for this well-known spectral model is cited in many ocean engineering references and is given by:

$$S(f) = \frac{0.81 \times 10^{-3} g^2}{(2\pi)^4 f^5} e^{-0.74 (g/2\pi U f)^4} \qquad 6.3$$

where f=frequency, g=gravitational constant, and U is the wind speed.

Figure 6.4 illustrates one example of a programmed and the corresponding measured (after subjective corrections) spectral functions taken from the Seakeeping Basin at the U. S. Naval Academy. Note that the original spectra have been converted to the corresponding amplitudes to better compare with HPT estimates.

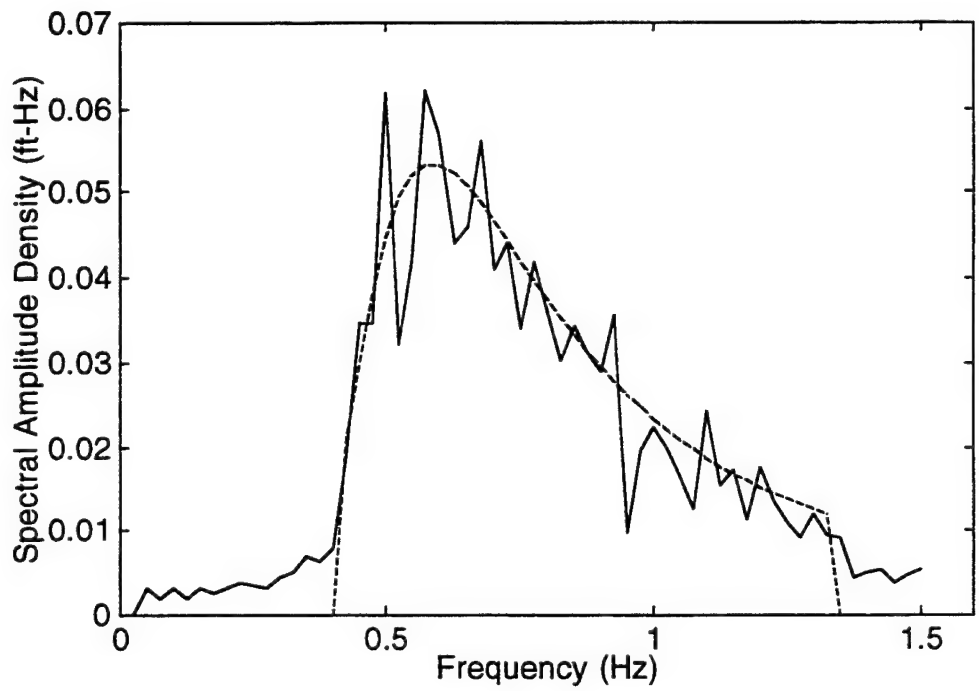

Figure 6.4 Programmed (- - -) and Measured (———) Amplitude Functions for U. S. N. A. Pierson Moskowitz Wave Signal

The HPT analysis began by high-pass (anti-aliasing) filtering the data and subsequently decimating by 15. Two parametric HPT analyses were performed using the following parameter sets:

total number of data points in signal:	410
number of points per HPT analysis:	120, 160
number of points between records:	30, 40
starting data point:	40
HPT reference FFT length for displays:	64

Both analyses used a 75 percent overlap, for a total of 7 and 5 analyses, respectively. The purpose of conducting two HPT analyses was to vary the frequency resolution to observe whether this changes the results.

The measured wave signal is shown in Figure 6.5.

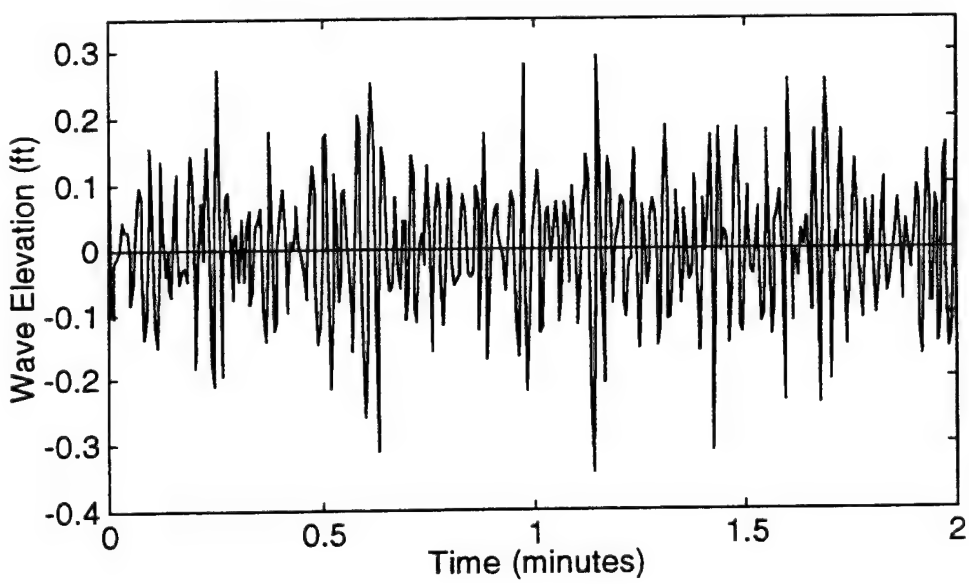

Figure 6.5 Measured Pierson Moskowitz Wave Signal

Figure 6.6 repeats Figure 6.4 except that it details the amplitudes in the neighborhood of the spectral peak and includes estimates from the shorter of the two HPT analyses.

If the signal was truly stationary (e.g., no reflections from the other end of the tank), then HPT should converge to the same frequency vector for each of the analysis. Inspection of Figure 6.6 shows that this did not happen. The next question is, is this uncertainty from HPT or is it due to nonstationarity in the signal?

One means of answering this is to conduct a parametric HPT study with different bin resolutions. Results from such a study are shown in Figure 6.7 in bin space for approximately the same frequency span as Figure 6.6. The lines have been added subjectively to aid the eye in tracing the evolution of the frequencies (and amplitudes) from the two analyses versus time. The HPT segment lengths were chosen to insure that the bin resolution was greater than the 0.025 Hz spacing used for the wavemaker frequencies (dotted lines); the corresponding bin resolution relative to Figure 6.7 is approximately 0.5.

In general, the two HPT patterns are similar, implying that the variability is inherent in the signal rather than being an artifact of the technique. Additional HPT information, for example from the phase continuity plots (not presented here), support this conclusion that the signal is not

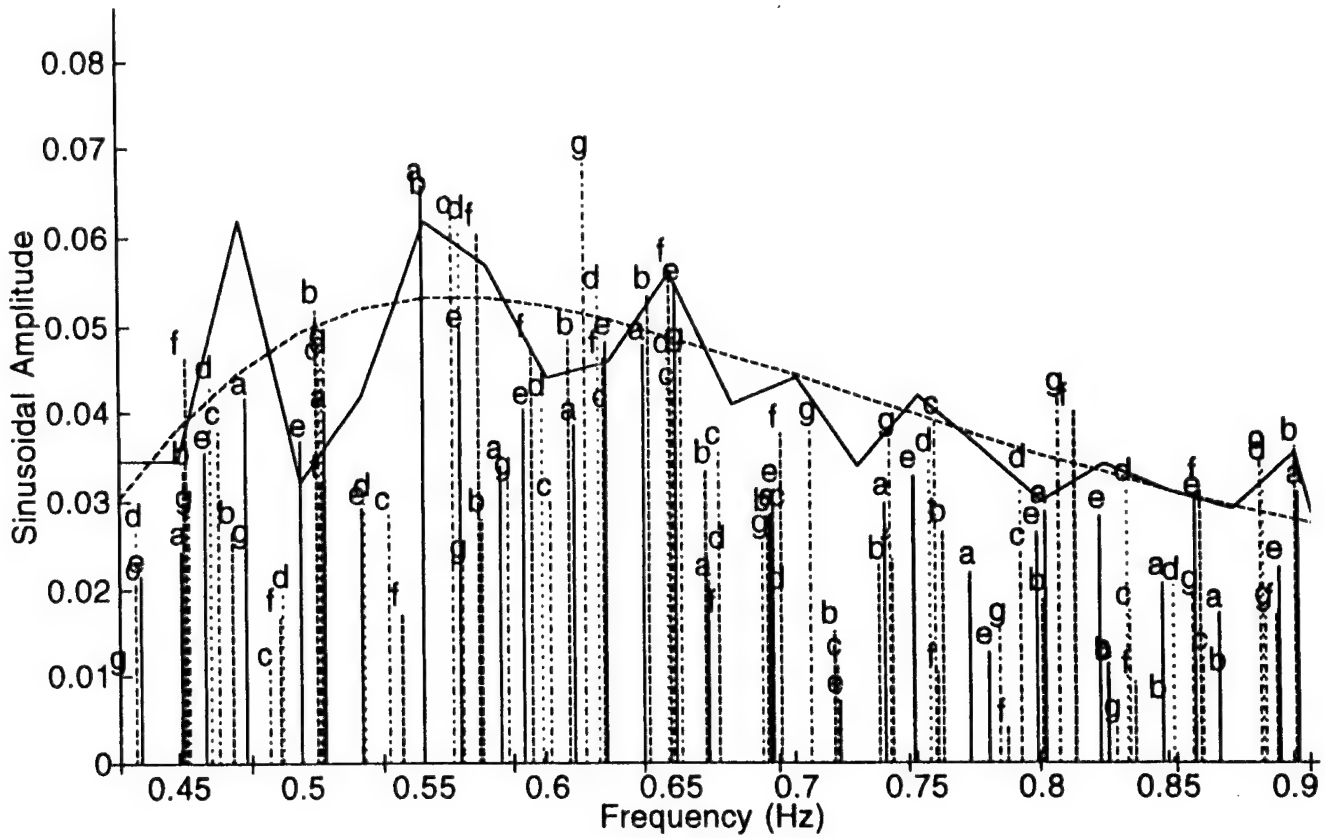

Figure 6.6 Sinusoidal Amplitudes for the Pierson Moskowitz Wave Signal; - - - = model; ___ = measured; a .... g = sequential HPT estimates

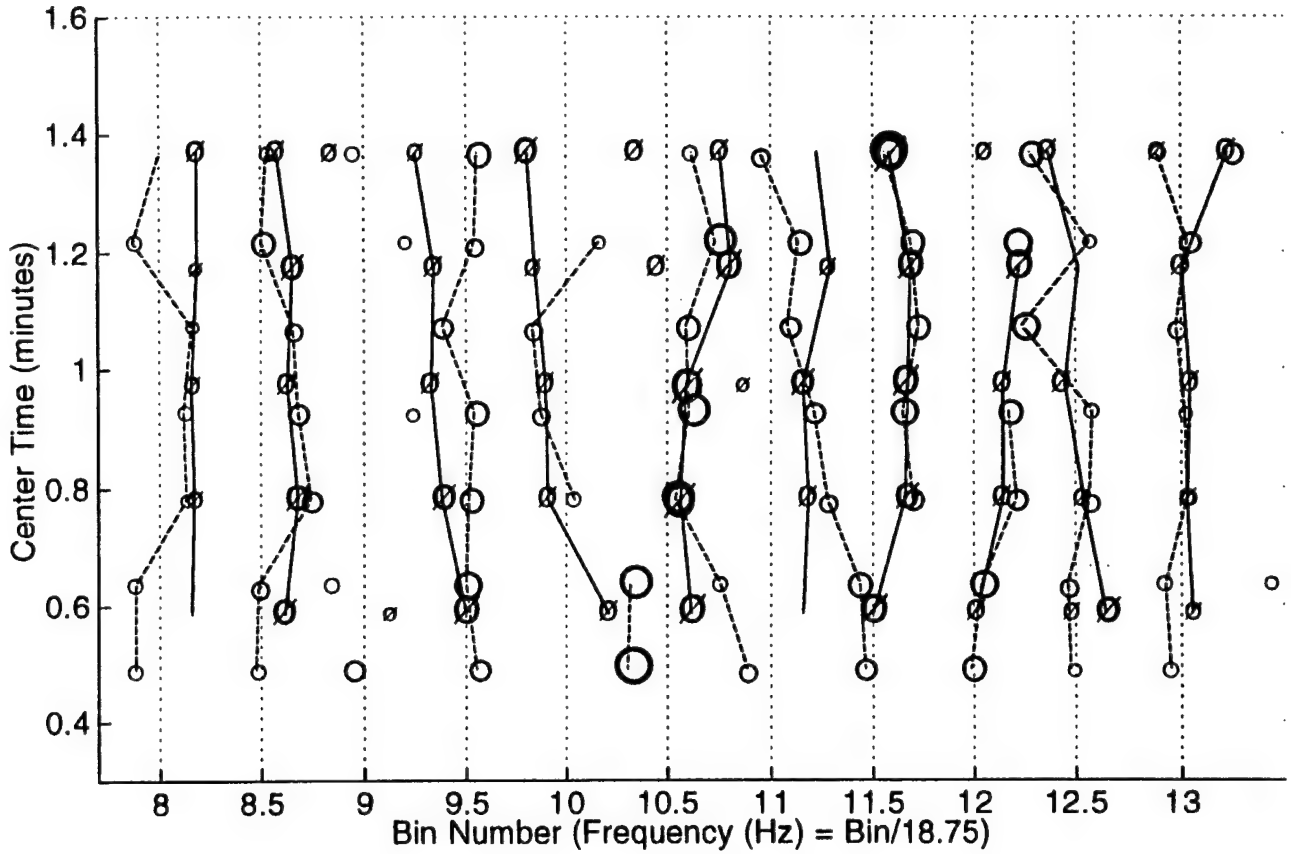

Figure 6.7 Comparison of HPT Component Evolution for the Pierson Moskowitz Wave Signal;

- - - = 120 point segment; ___ = 160 point segment

comprised of sinusoids with equally-spaced (Fourier) frequencies. One plausible cause for this uneven frequency spacing is internal resonances in the hydraulics for the wavemaker. This could also explain why it is typically necessary to subjectively adjust laboratory wave generators to get the desired spectrum, namely, that the actual component frequencies are not orthogonal and therefore leak finite energy to neighboring frequencies. While one study such as this is not definitive enough to draw reliable conclusions, it once again demonstrates the potential of HPT information for real-world engineering applications.

6.5. Frigate Heave and Wave Signals

This last section introduces a new conceptual argument that reinforces the use of HPT for interpreting signal characteristics. This example involves two correlated laboratory signals measured at the U. S. Naval Academy: a broadband wave and the associated heave response of a model frigate (FFG) in beam seas. Details are reported in Zseleczky and Wallendorf, 1994. The data is in fact collected from two separate tests. The first measures waves with the FFG out of the water, while the second uses the same pattern of waves from the wavemaker with the model present; this insures that the wave measurement is not contaminated with radiated waves from the model. (The wavemaker has been proven to reliably and repeatably produce identical time series if the same program is used.)

The HPT analyses used the following parameters:

total number of data points in signal:	600
number of points per HPT analysis:	200
comparable FFT length:	256
number of points per shift:	50
starting data point:	60
HPT reference FFT length for display:	128
sampling period (sec):	0.333

Example amplitude versus bin number results from HPT are shown in Figure 6.8 for one segment of the waves (excitation) and FFG heave (response). Because a 128 point FFT length was used as the reference for these HPT results, the programmed components based on a 256 point FFT would be expected at half bin spacings (i.e., twice the inherent resolution of a 128 point FFT).

Figure 6.9 illustrates HPT component evolutions of both signals over a representative bin number range. The significance of Figures 6.8 and 6.9 is how remarkably similar and invariant corresponding bin numbers are over the entire band. *This provides a strong argument that the HPT estimated frequency vectors are correct since this is a linear system, and it is a property of linear systems that linear superposition holds and the component frequencies are invariant.*

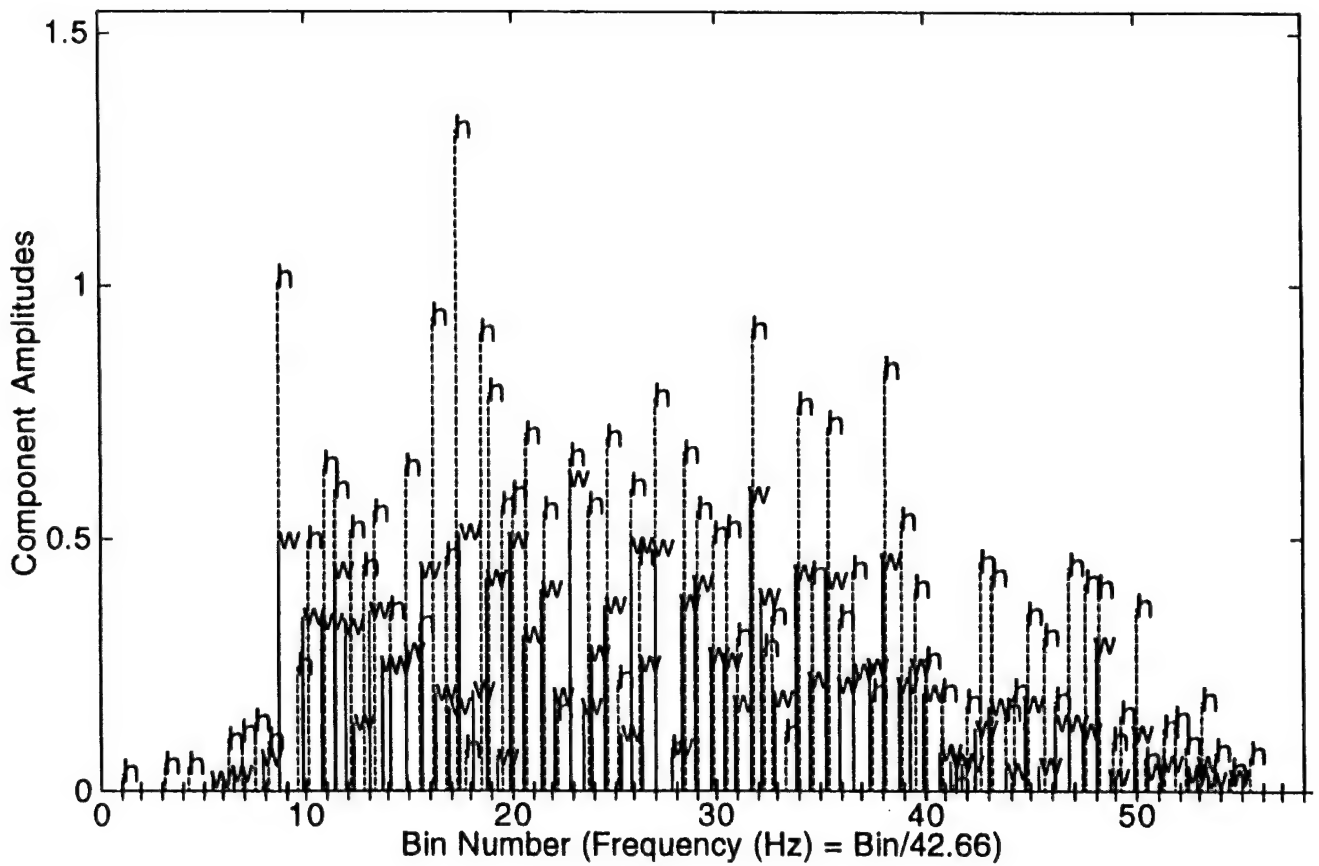

Figure 6.8 Comparison of Amplitudes versus Bin Number for Wave (---) and FFG Heave (- - -) Signals Over Full Bandwidth

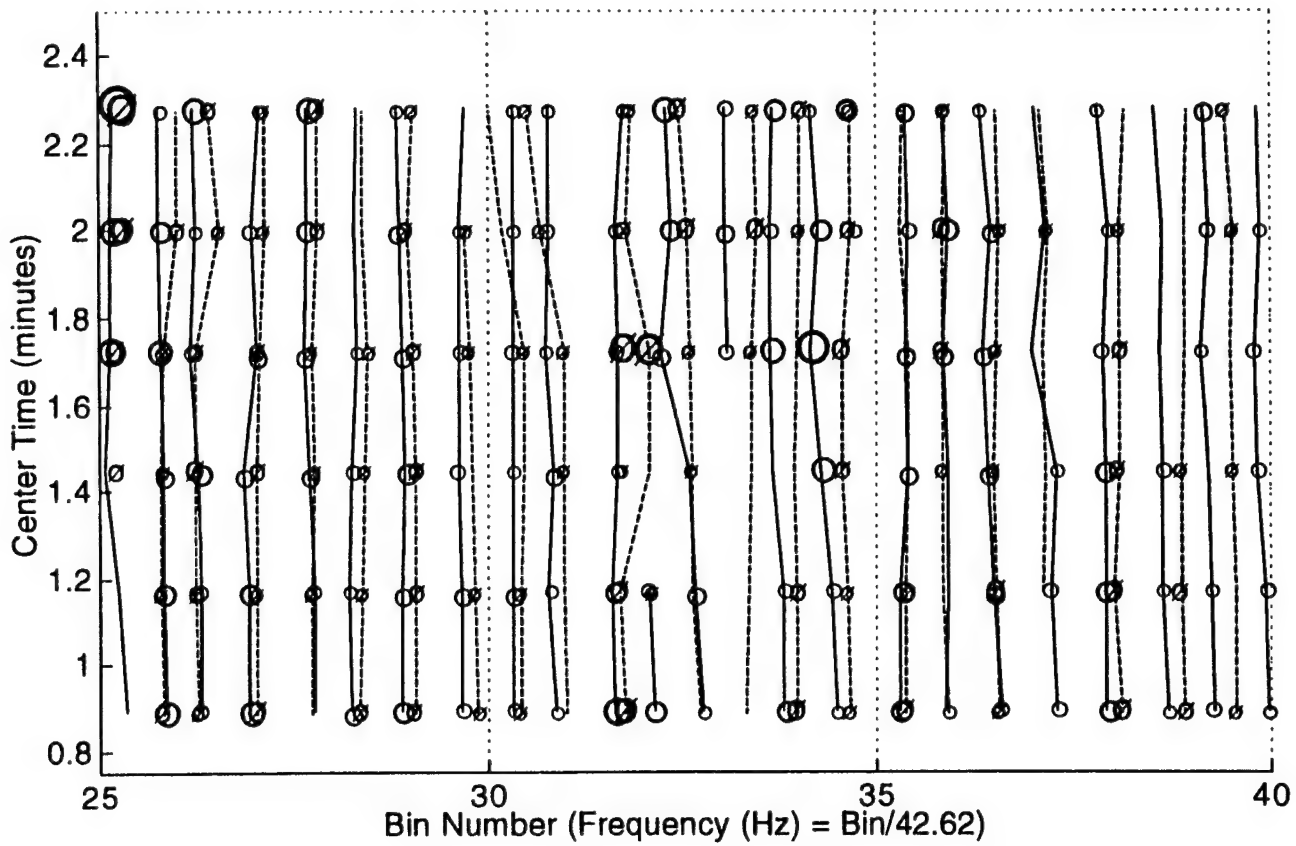

Figure 6.9 Comparison of HPT Component Evolution for the Wave (----) and FFG Heave (- - -) Signals

6.6 Chapter Summary.

This Chapter has demonstrated the use of Harmonic Phase Tracking (HPT) using a variety of physical signals. In each case the HPT results added significant information not available from either harmonic analysis or traditional spectral analysis. While it is true that some of the high resolution signal processing techniques such as MUSIC could have been applied to the tide and rank 4 wave signals, the Pierson Moskowitz and FFG signals had much higher rank and are not good candidates for those techniques.

The capability of HPT to identify and track individual wave components has many engineering applications. This Chapter has demonstrated only a few of those applications with respect to ocean engineering, such as detecting the arrival of a reflected wave in a laboratory basin or actual frequency vector from a wave generator, .

Also, HPT was used in the FFG study to estimate the frequency response function of the wave:heave system. Those results were essentially comparable to standard FFT-based estimates using cross spectra and are not shown. While the examples in this and the preceding Chapter have demonstrated the large bias inherent in FFT-based *absolute* phase estimates, the *relative* difference between these phases is typically correct. Given that the FFT *consistently* biases the absolute phases this

conclusion is perhaps not that surprising. Accordingly, it is concluded that the use of HPT does not provide any additional information compared to the tremendously less computationally-intensive spectral methods when system functions are under study. On the other hand, if HPT is used to investigate other signal characteristics, then as shown in Chapter 7 it does provide reliable estimates of the magnitude and phase of frequency response functions.

Chapter 5 demonstrated that HPT is capable of handing all of the characteristics expected in stationary and slowly-varying multiharmonic signals. This Chapter continued that investigation by showing that the technique readily modeled forced (i.e., directly generated by astronomical or physical means) wave signals. The foundation is now complete for the study of free ocean waves in Chapter 7.

[blank]

CHAPTER 7

HARMONIC PHASE TRACKING ANALYSIS OF OCEAN WAVE FIELDS

7.1 Chapter Introduction

This Chapter uses Harmonic Phase Tracking (HPT) to describe ocean waves in an entirely new way. Both temporal and spatial investigations are presented that, in a manner of speaking, allow for "locally deterministic" interpretations of the ocean surface rather than the less informative "sum of an infinite number of sinusoids with infinitesimal amplitudes arriving from a continuum of directions" or "sum of orthogonal sinusoidal basis functions".

As described in Chapter 2, traditional ocean wave analysis assumes that since the signal is stochastic with unknown characteristics, the choice of analysis tool must be general enough to accomodate any conceivable characteristic. As a result, orthogonal methods such as Fourier Series and

most recently the closely-related wavelet techniques are commonly used. Both of these are low resolution techniques; for example, the spectrum is only capable of estimating averaged quantities over frequency and time. So, while this is a safe approach that does describe the entire wavefield, the disadvantage is that the information it provides does not provide much physical insight into the local behavior of individual waves or wave packets.

The objective of this Chapter is to demonstrate how HPT can more usefully quantify local wave fields in time and space. Two representative sets of wave data are used. As described in Section 7.2, all of the wave measurements were provided by the Army Corps of Engineers Field Research Facility at Duck NC. The first set of waves was selected to represent (quasi-) stationary conditions and is presented in Section 7.2. The second highly nonstationary wave set corresponds to Hurricane Bob in 1991. This is described in Section 7.3 and is the focus of this Chapter. Chapter conclusions are summarized in Section 7.4.

7.2 Description of FRF Wave Data

The waves were selected from measurements made by the Army at their Field Research Facility (FRF) at Duck NC. The general area is described by Birkemeier, et. al., 1985. Essentially, this section of coast has a straight coastline and a very gradual increase in water depth parallel to the shore.

The data used in this study comes from an orthogonal array of pressure sensors located 1 km offshore and bottom mounted in 8 meters of water. As stated in Chapter 2, this depth is sufficient to avoid strong nonlinearities in the wave surface and the corresponding coupling in the frequency domain components. The long (predominantly North-South) axis of the array is parallel to the 8m depth contour.

Figure 7.1 shows the general arrangement of the gages used in this study. The numbers refer to the gage identifier. The coordinates relative to the local origin (Gage 131) are given in Tables 7.1a and 7.1b. The North-South axis descriptor refers to the long axis parallel to the 8m contour, and the East-West descriptor is then the on-shore orthogonal axis.

Gage Number	North-South Coordinate (m)	East-West Coordinate (m)
191	189.9	-0.2
181	155.3	-0.2
171	130.0	-0.5
111	25.0	-0.1
121	15.3	-0.1
131	0	0
151	-39.9	-0.2
161	-64.9	-0.3

Table 7.1a. Coordinates for North-South Array FRF Gages

Gage Number	North-South Coordinate (m)	East-West Coordinate (m)
211	-0.1	-79.9
221	0.1	-39.8
231	0.1	-9.8
131	0	0
241	0.2	20.3
251	0.1	40.2

Table 7.1b. Coordinates for East-West Array FRF Gages

Note that the maximum gage separations are 255m and 120m in the North-South and East-West directions, respectively. Second, it is noted that data from Gage 221 was used in the quasi-stationary HPT analysis in the next section, but this gage was not working for the hurricane waves and Gage 231 was used instead. Regardless, each analysis had data from 12 gages available.

The data was sampled at 0.25 second intervals for the quasi-stationary data and 0.5 second intervals for the hurricane data. The pressure data was digitally recorded at the FRF for approximately 2-2/3 hours over 3 hour intervals, so it is essentially continuous for the purposes of this study. The pressure was low-pass filtered and converted to instantaneous amplitude at the FRF prior to distribution.

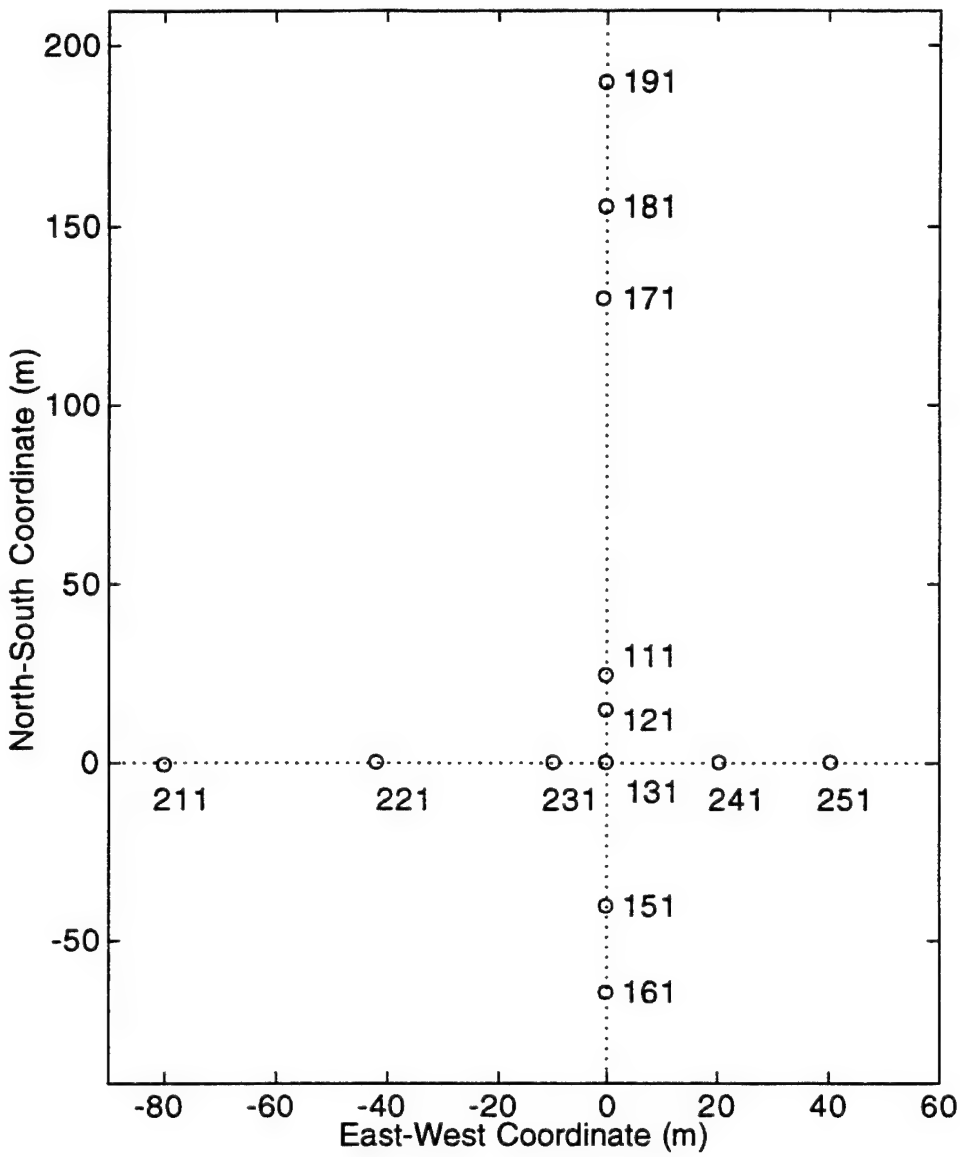

Figure 7.1 Gage Identifiers and Coordinates for FRF 8 meter Array

7.3 Quasi-Stationary Wave Field Analysis

It was inferred from inspection of monthly summary statistics reports provided by the FRF that the waves on the morning of September 13 1990 were very stationary, at least as measured by the significant wave height and peak period averaged over 3 hour intervals. Supportive environmental statistics are presented in Table 7.2. In addition to the observations in Table 7.2, visual and radar estimates of the wave direction were consistent at 90 to 95 degrees on the mornings of 12, 13 and 14 September. Waves starting at 0400 on September 13 were chosen for detailed analysis.

	Wave Descriptors		Wind Descriptors	
Time (hrs)	RMS (m)	T_{peak} (sec)	Vel (m/sec)	Dir (deg)
0100	0.24	12.2	4	52
0700	0.23	12.2	6	50
1300	0.24	12.2	5	48

Table 7.2 Environmental Parameters for Quasi-Stationary Waves

The FRF data was prepared prior to the HPT analyses to emphasize the main bandwidth. Both low and high pass filters were used to minimize low

amplitude "noise", and the data was then decimated to yield a time step of 2 seconds. A sample of the waves is shown in Figure 7.2. Spectra for Gage 131 based on hourly averages are shown in Figure 7.3, representing 5 hours of wave data. 256-point FFTs were used, with 3 degrees of freedom (DOF) per spectrum. These five spectra are reasonably stationary in form, with rms values of: 0.170, 0.171, 0.173, 0.161 and 0.172 m, respectively; peak periods are: 11.59, 11.65, 11.32, 11.59, and 11.65 sec, respectively.

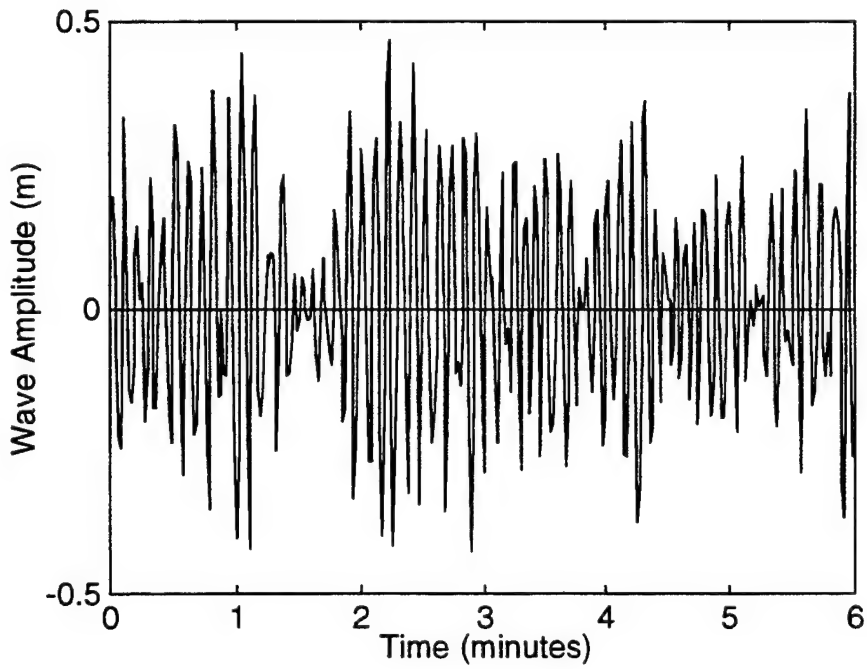

Figure 7.2 Sample Wave Data for Quasi-Stationary Analysis

Note that this wave field is unimodal and very narrowbanded (which is why the data could be low and high pass filtered with no loss of significant information).

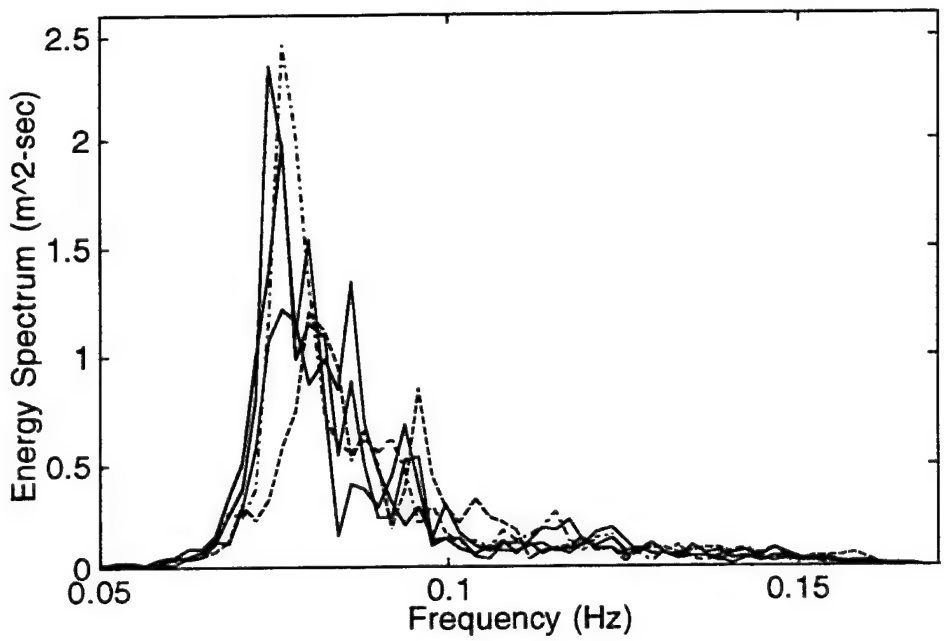

Figure 7.3 Spectra for Quasi-Stationary Wave Data; solid lines = 0400-0650, dashed line = 0700-0755, dot-dash line = 0755-0850.

The only step associated with using HPT is specifying the segment length. The default segment length for all of the wave analyses in this Chapter was 300 points = 600 seconds = 10 minutes. This was consistent with recommendations for spectra in Tucker (1991) and is just shorter than the 20 minute segments recommended by Goda (1985). However, since HPT uses a 25 percent extension before and after this 300 point segment, it effectively uses 450 points (or 15 minutes) per HPT analysis. Therefore, 512-point segments were defined as having equivalent resolution for any complementary spectral calculations.

One final observation is made regarding comparisons between HPT and FFT-based frequency domain mappings. Since HPT allows for arbitrary frequencies for each segment, simple averaging of "similar" components among the segments is not as well defined as with the FFT and/or spectrum. This is not necessarily an advantage of the FFT; recall how poorly a FFT models nonstationary signals with non constant frequencies as presented in Chapter 5. Thus, in some instances it will be more consistent to compare HPT results to raw (i.e., based on *one* FFT) rather than ensemble-averaged spectral functions to best appreciate the differences between the two techniques. This is done with full understanding that the resulting Fourier transforms are statistically unreliable with 100 percent error bounds. The choice of using raw or averaged spectral functions and the consequences will be carefully expressed whenever either is used.

For example, consider calculating raw 512-point spectra for many of the gages over the same time interval. (This is in fact the correct interpretation of ensemble averaging if the wave field was indeed spatially stationary.) This was done using wave records near the beginning of the 0400 block. Gages 111, 251, 161, and 211 were selected based on the approximately-equal separations of 100 meters. The energy spectra were converted to amplitude spectra and are shown in Figure 7.4a.

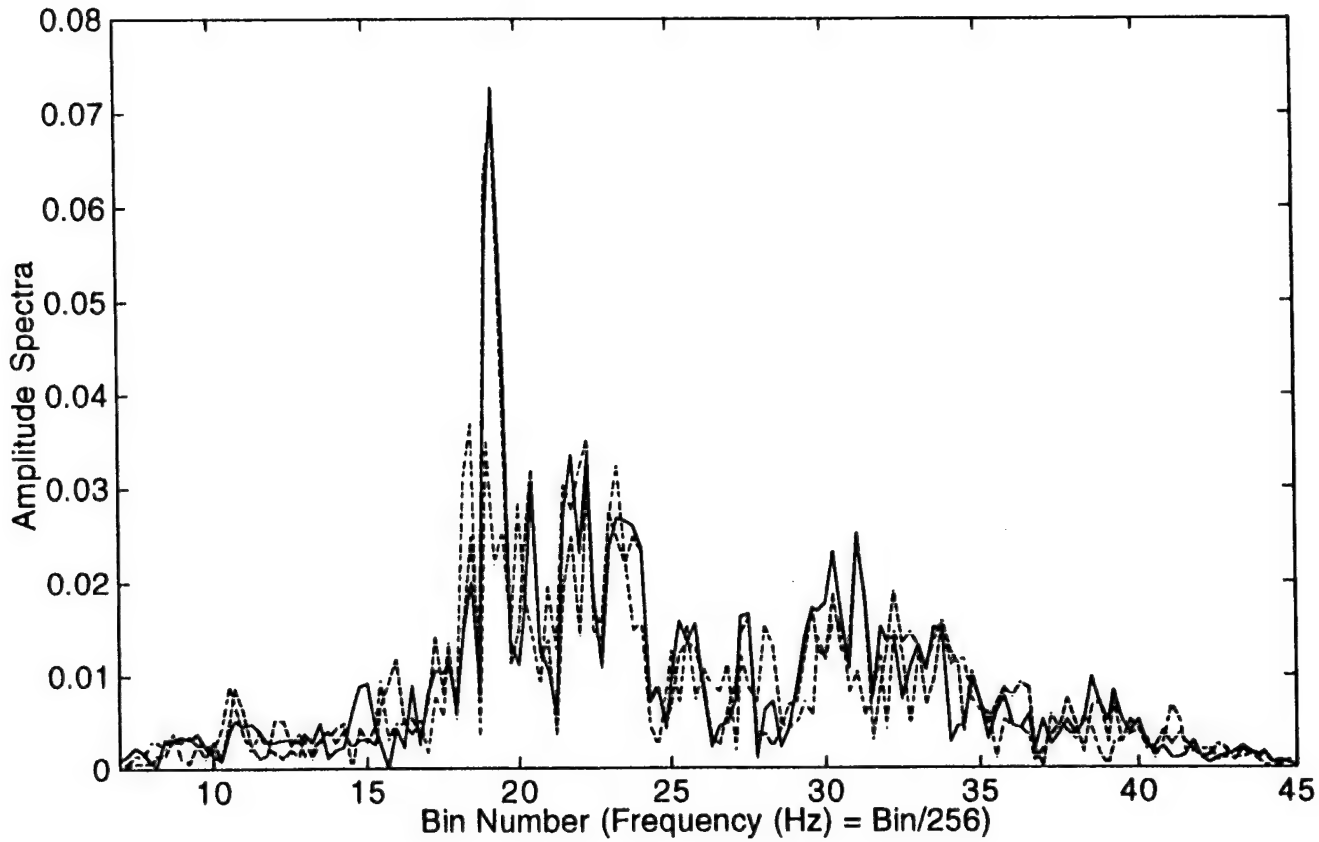

Figure 7.4a. Raw FFT-based Amplitude Spectra for 0415-0425, September 13 1990; ——— = Gage 191; - - - = Gage 251; – · – · – = Gage 161; and . . . . . = Gage 211.

They are reasonably similar considering the separation between the gages and the lack of averaging.

The same functions near the spectral peak are included in Figure 7.4b, superimposed with HPT amplitude estimates for the same gages and the same time period. There are several important conclusions from this figure. First, observe the consistency of most of the HPT discrete frequency estimates over the entire band. Second, the HPT amplitude estimates are likewise generally consistent among the four gages. And third, and perhaps most important, observe how the general peaks from the two techniques are not particularly consistent; the most obvious difference is just above bin 22 where HPT shows a large component with an amplitude that is comparable to the spectral peak which is not detected by the [raw] FFT estimate. The same is true between bins 16 and 18, and at bins 20.5 and 21.3 where the large harmonics are not at integer bin numbers and are hence averaged by the FFT into neighboring bins.

Given these differences, just how accurate are these HPT estimates? Are they representative of the actual components (packets) in the wave field, or are they numerical artifacts? And if they are true, are the FFT estimates that biased? Answering this is addressed two ways. (Analytical expressions for HPT bias and variance were not found in this study; that topic is addressed in the next Chapter.)

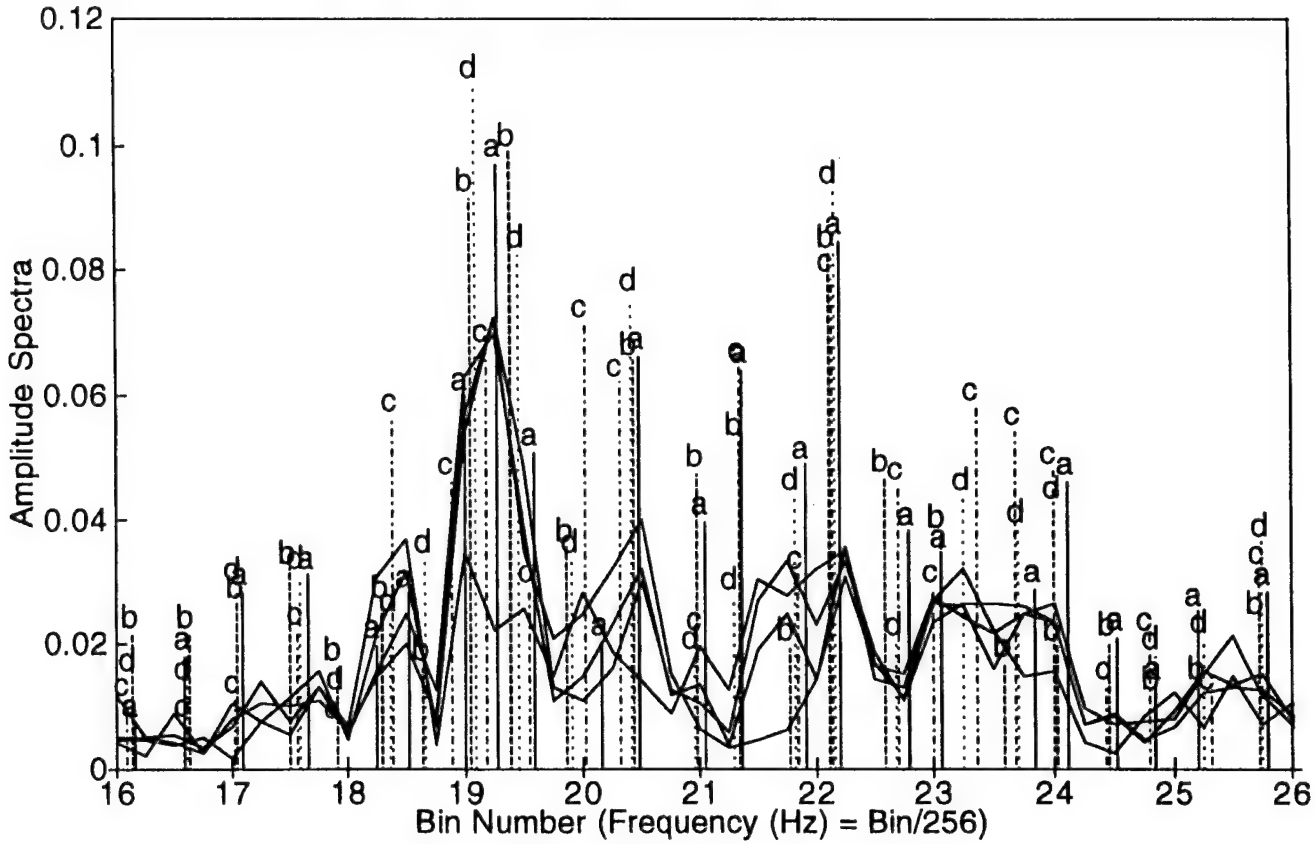

Figure 7.4b. Representative HPT Components for 0415-0425, September 13 1990; a = Gage 191; b = 251; c = 161; and d = 211. Lines are Raw FFT-based spectra from Figure 7.5a.

First, recall that Chapter 5 did present a brief investigation demonstrating that HPT results were fairly insensitive to the initial starting vector; one additional example is presented here that confirms that conclusion for ocean waves. Figure 7.5 details two independent HPT estimates around the spectral peak for Gage 211. The one labeled "d" was independently calculated directly from the time series using the procedures in Chapter 4. The other estimate labeled "c" used the converged HPT estimate from Gage 161 for the initial frequency vector. Observe that the two estimates are strikingly similar in frequency and amplitude. This is the first evidence that HPT results are at least consistent for ocean waves.

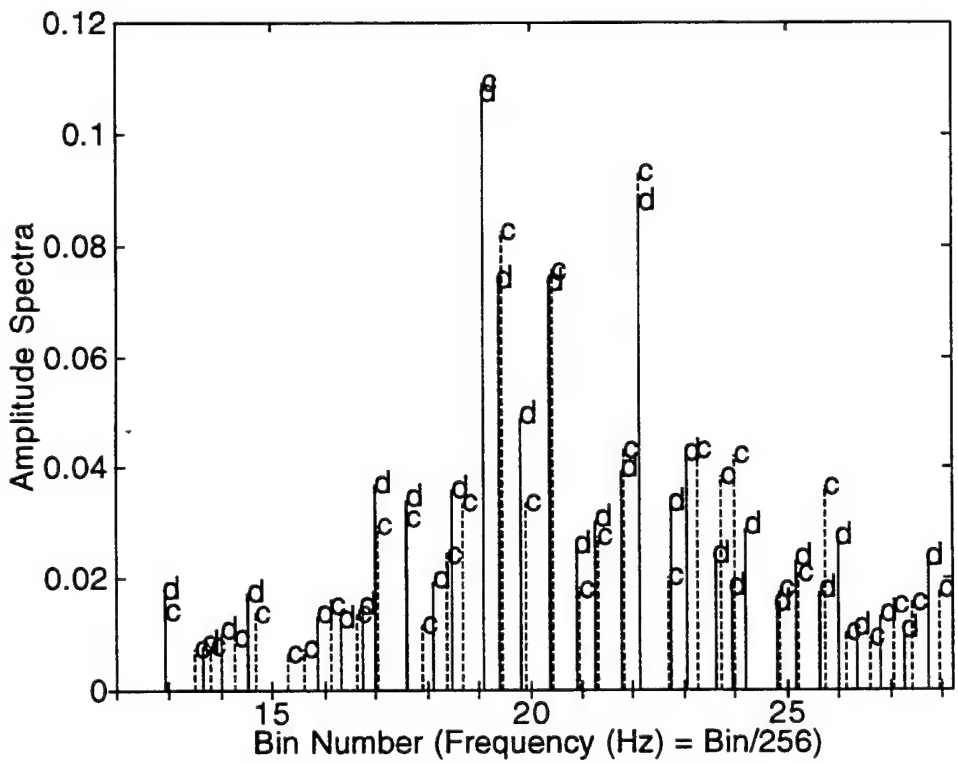

Figure 7.5 Consistency of Direct and Correlated HPT Estimates for Gage 211

The second and better method for assessing the accuracy of HPT results is to analyze a series of overlapping time series segments and inspect the component parameter evolution. This approach was illustrated numerous times in the last two Chapters and is particularly informative in cases like this where the signal is known to be stochastic. Inspection of the evolutionary results allows for "smoothing" of frequencies and/or amplitudes in time according to the observed trends, thereby defining individual packets with "averaged" rather than "raw" parameters.

Figure 7.6 shows the evolution of HPT components for Gage 131 for this stationary wavefield. A shift of 2 minutes, or 20 percent, was used relative to the 10 minute analysis segments (i.e., 80 percent overlap). The first observation is that there appear to be definite, discrete wave "packets" with slowly-varying frequency and amplitude. This is an encouraging conclusion since the HPT model assumes a finite summation of discrete harmonics to model the signal. This is also the first confirmation by HPT widely-held suspicions that the ocean does self-organize into discrete packets that persist for time scales of minutes or longer.

The second observation is that almost all of the frequencies of these packets steadily decrease in time, many at a surprisingly high rate. This may or may not indicate true long-term downshifting of energy to lower frequencies. Recall that Chapter 5 presented one analytical example of

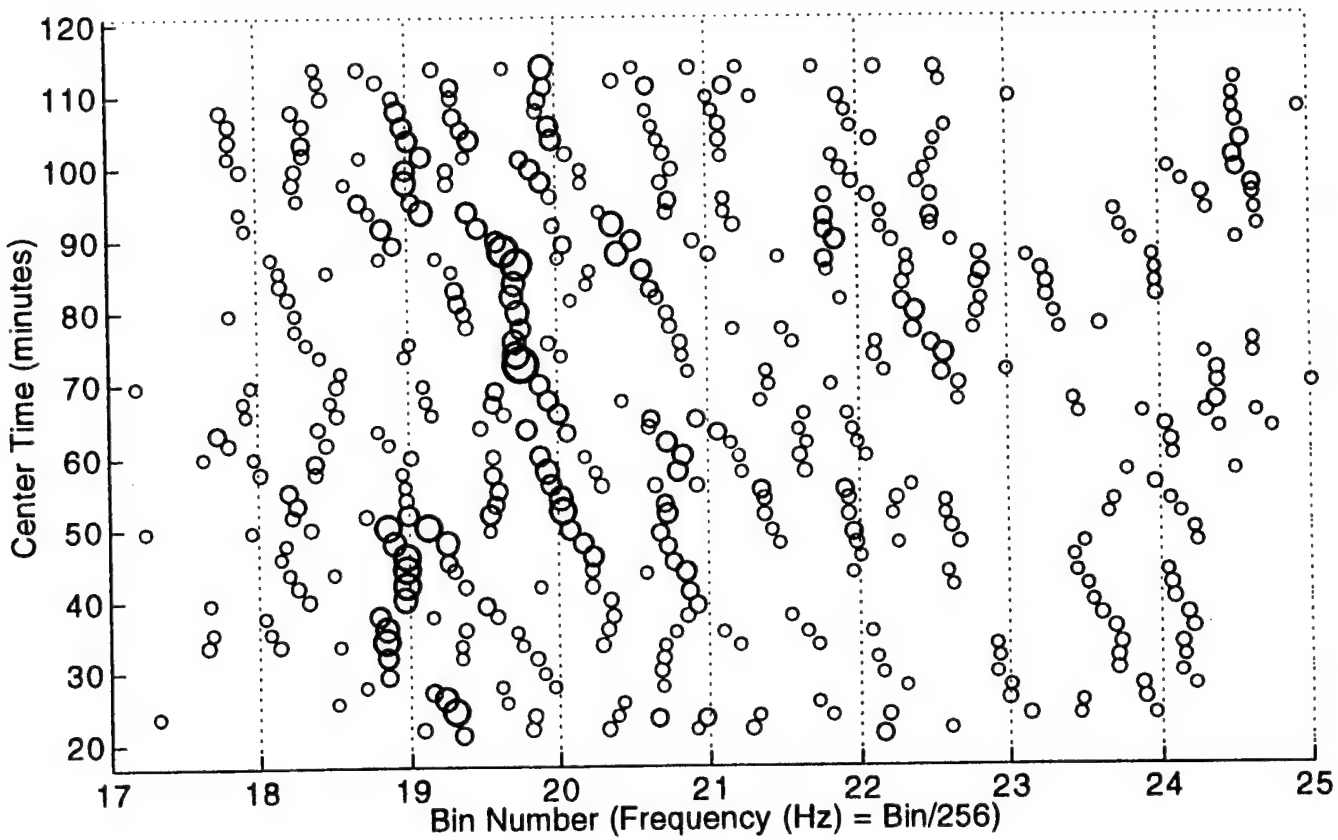

Figure 7.6 HPT Evolution for Gage 131, 0415-0615, September 13, 1990.

three stationary harmonics where the instantaneous frequency, and hence the HPT estimated frequency, showed an apparently nonstationary "sawtooth" type behavior defined by a steadily-decreasing frequency over much of the wave cycle followed by a ill-defined rise to begin the steady downshift again. This was illustrated in Figure 5.19. In some respects, the patterns in Figure 7.6 for these ocean waves are similar to the analytical patterns, which leads to a preliminary conclusion that the phenomena are related. However, there is also a fundamental and important difference between the ocean and analytical component evolutions - namely, that the ocean packets show a large degree of overlap relative to the steady downshifting intervals that is not present in the analytical counterpart.

The concept of HPT phase continuity introduced in the last two Chapters can be used to great advantage here to reliably determine whether these HPT-identified packets truly represent physically continuous wave forms. The definition of a packet is open to debate; for the purposes of this investigation, a packet ends when the bin number changes by more than half of the bin resolution of the associated HPT analysis - (this partly explains the reason for the high degree of overlap between adjacent HPT analyses - namely, to track nonstationary frequency shifts) or when the amplitude is negligible in the next HPT estimate. Frequencies, amplitudes and phases for two representative packets are presented in Figures 7.7 and 7.8. The first packet evolution in Figure 7.7a has a relatively constant frequency versus time and accordingly is labeled "stationary". This

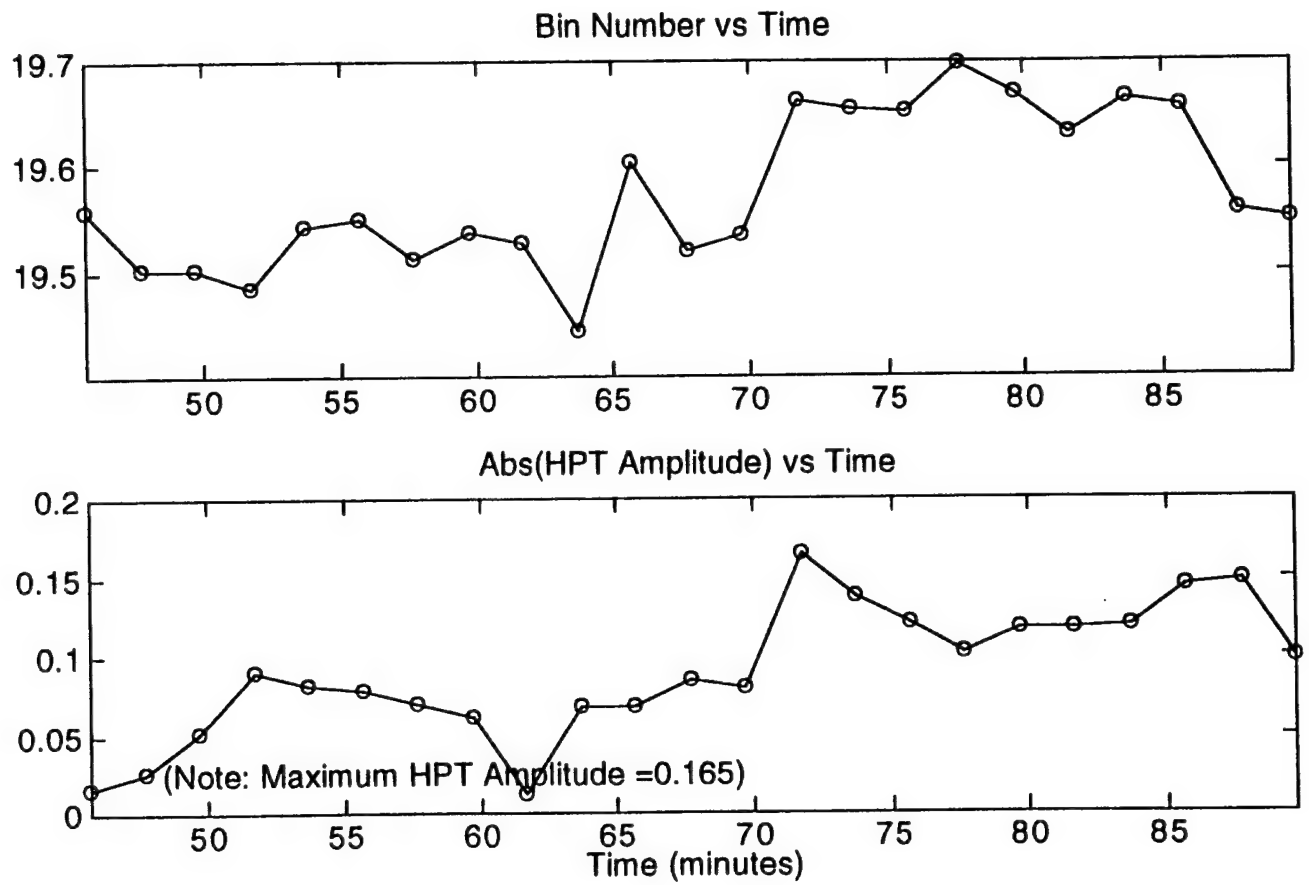

Figure 7.7a Frequency and Amplitude Evolution for a Representative Stationary Packet

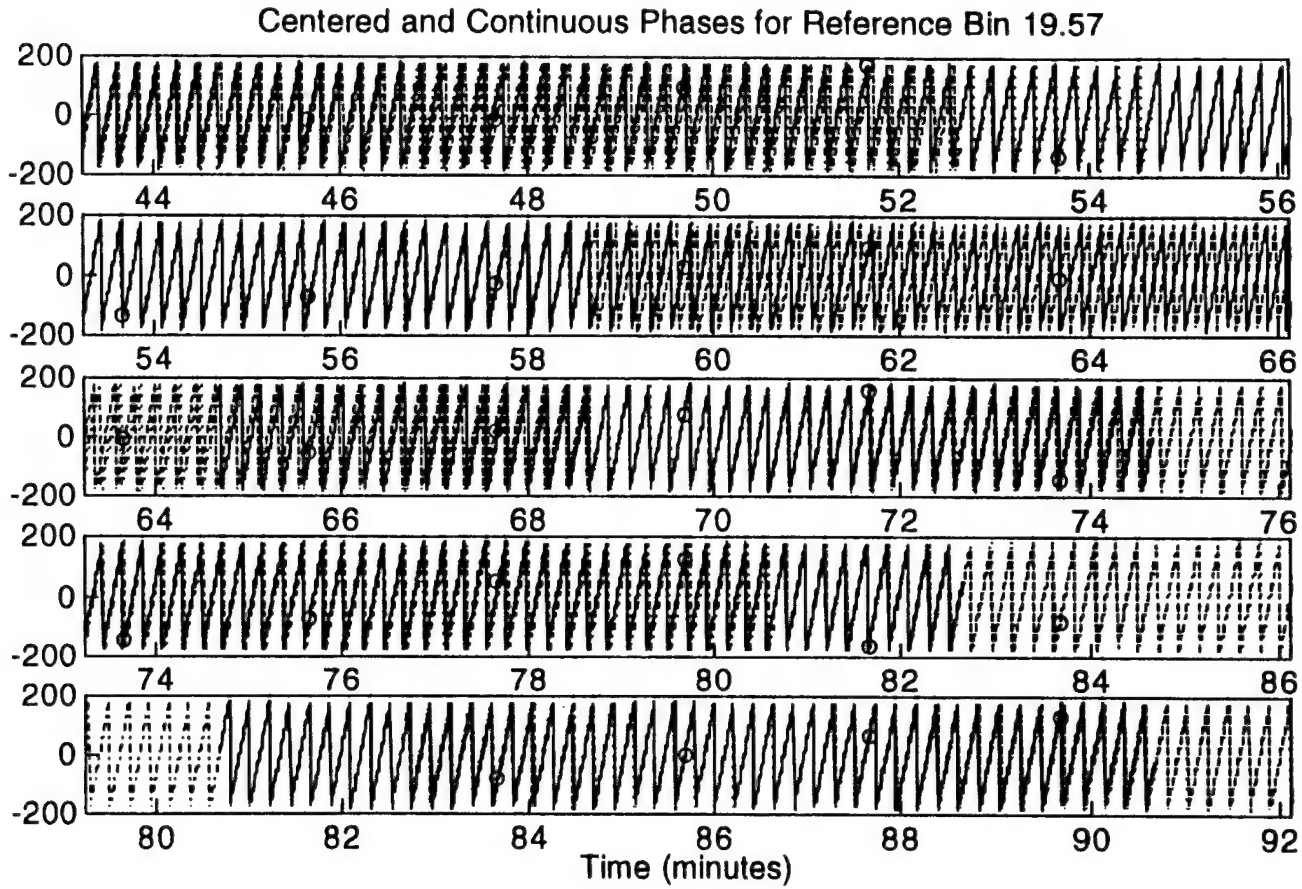

Figure 7.7b Phase Evolution for a Representative Stationary Packet

particular packet is most easily correlated with Figure 7.6 using the frequency information in the upper figure. But although the information in Figure 7.7a is relatively well-behaved, it cannot be considered definitive in terms of proving that a physical packet exists; for that we need to inspect the phase function in Figure 7.7b. Together, these figures show that this subject packet was generally continuous, with some phase and frequency distortions evident between 60-67 minutes (recall that these HPT analyses used a 10 minute segment that does introduce inherent averaging over that time scale). Interestingly, according to Figure 7.6 this is the time when a second packet from a higher frequency merges into the subject packet. A second packet merges, again from a higher frequency, just after 70 minutes. While this merger is not reflected by a phase distortion, it does result in a significant increase in the amplitude of the subject packet,. as might be expected. All of these observations lead to the tentative conclusion that HPT parameters may truly reflect physical processes in the wavefield.

The second packet example corresponds to a packet with a changing frequency; this is labeled the "nonstationary" packet and is plotted in Figures 7.8a and 7.8b. Correlations to Figure 7.6 will show that this is, in fact, the first packet that merged into the packet used in the discussion immediately above. Observe the steady, well-behaved decrease in the frequency and increase in amplitude. However, it is the phase function in Figure 7.8b that is the most informative. It shows two time spans with large phase distortions. The first span, between 25 and 32 minutes,

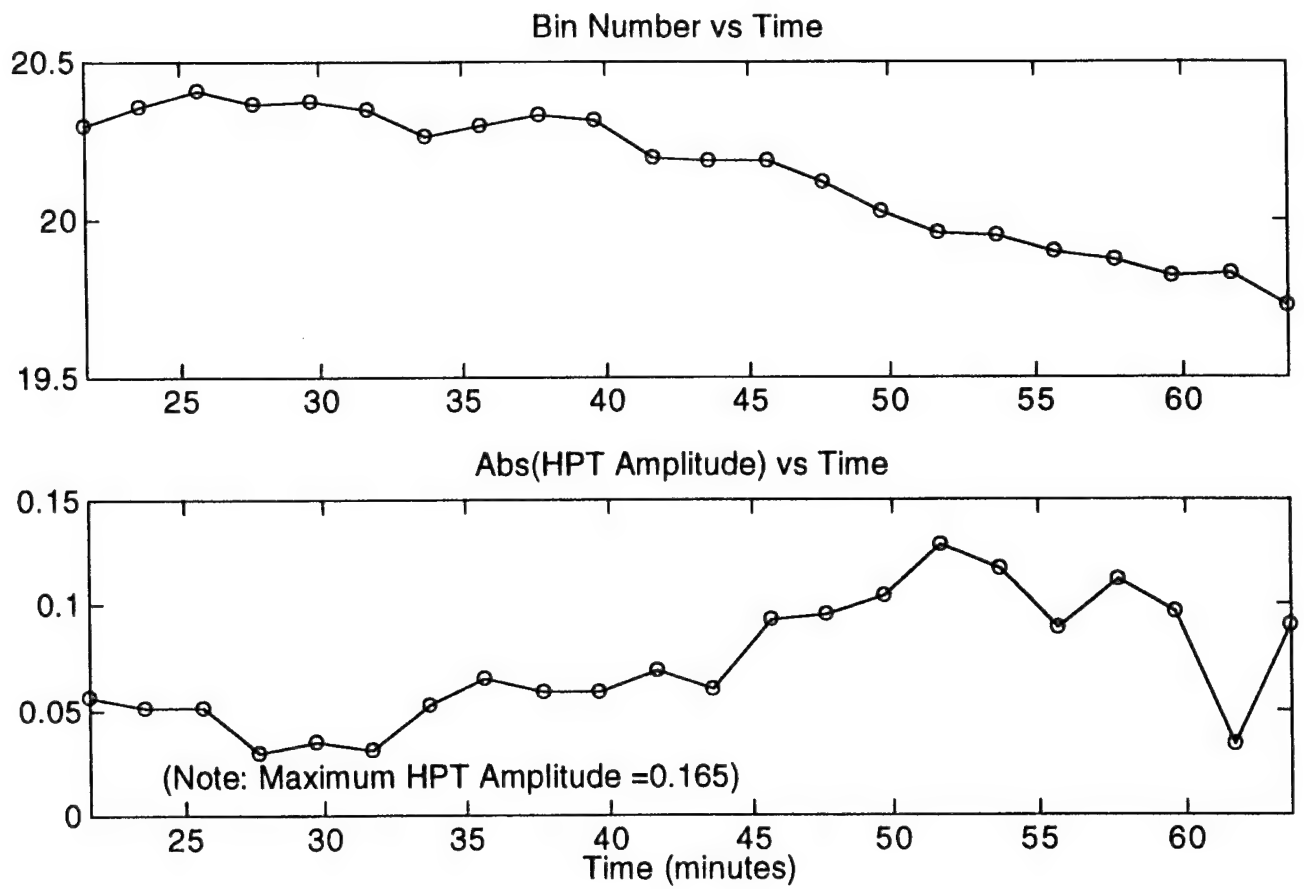

Figure 7.8a Frequency and Amplitude Evolution for a Representative Nonstationary Packet

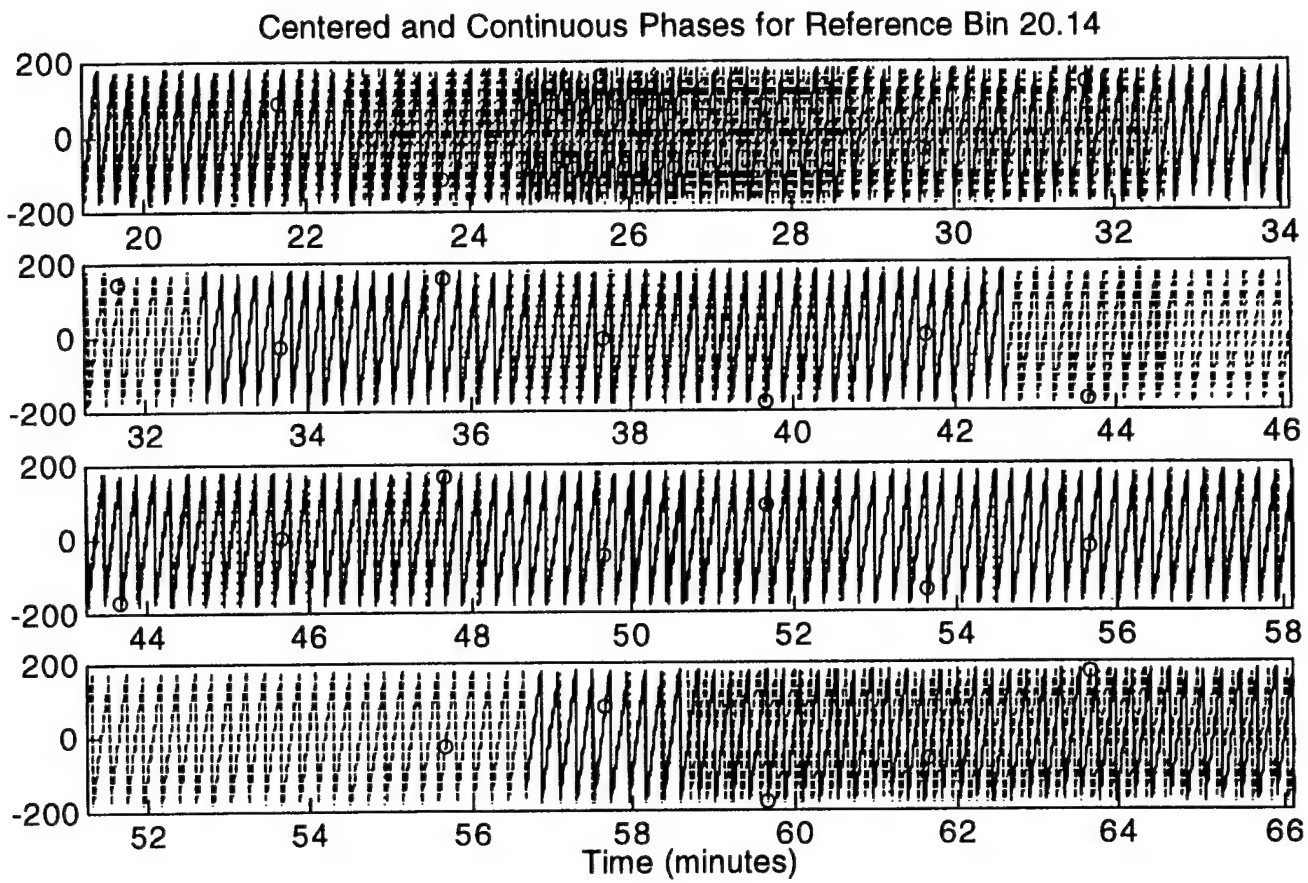

Figure 7.8b Phase Evolution for a Representative Nonstationary Packet

corresponds to very low amplitudes (Figure 7.8a), which illustrates how difficult it can be to correctly interpret these phase figures. The second time span with phase discontinuities is at the end, where this packet merges into the previous packet (or does the previous packet merge into this one?). As before, all of the HPT parameters for this packet provide information that is consistent with the behavior of adjacent packets. Inspection of the parameters for other packets was similar and is therefore not presented.

Based on these discussions, it is proposed that the HPT wave packet information shown in Figure 7.6 does model the wave field much better than the traditional spectral (or wavelet) model with respect to understanding the underlying physics. Since the focus of this wave study is the hurricane waves in the next section no further study of these stationary waves is presented. As given in Table 7.2 the rms value was only 0.25m, which means that the energy in the primary wave signal is so small that it could have been easily corrupted by waves from many incident directions.

7.4 Hurricane Bob Wave Field Analysis

This section is divided into three subsections. Subsection 7.4.1 describes the overall storm. Subsection 7.4.2 presents a variety of HPT estimates regarding packet characteristics, wavelengths, and incident direction versus segment length and selects an "optimum" time span for quantifying these waves. Subsection 7.4.3 contrasts various wave field descriptors for Hurricane Bob during the build-up, at the peak, and after the peak.

7.4.1 Overview of Storm

Hurricane Bob was a large storm that passed 40-48 km offshore of the FRF on August 18 and 19, 1991. The maximum wind speed exceeded 23 m/sec around midnight, with a maximum significant wave height of 4.8m. Further information regarding the storm is available from the FRF Preliminary Data Summary for August 1991 and other sources.

Figure 7.9 shows the significant wave height (defined as 4 times the rms value) at Gage 131 starting at 1900 on August 18 1991 for half hour averages. The solid lines connect estimates from contiguous data blocks of data; the dotted line is midnight.

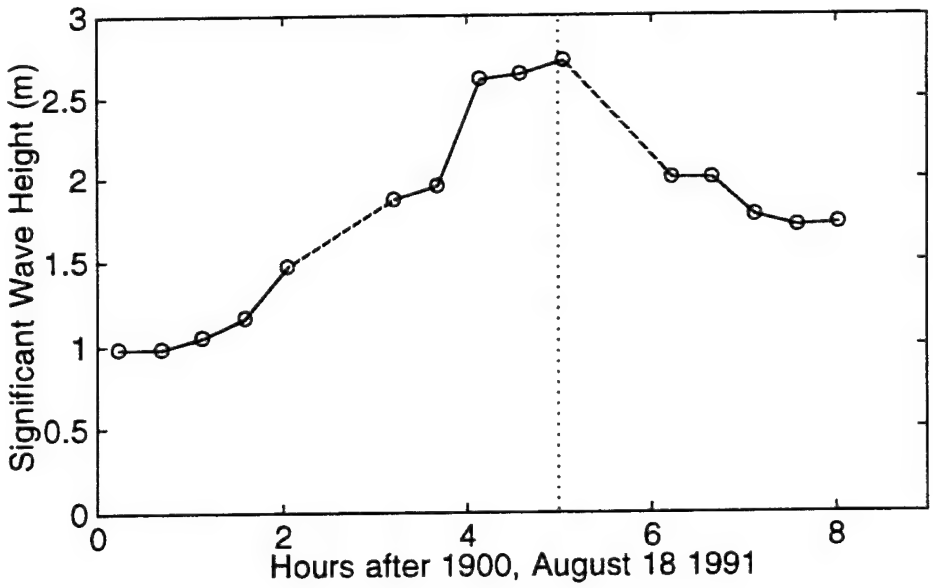

Figure 7.9 Significant Wave Height for Hurricane Bob

Corresponding spectra are shown in Figure 7.10 for "preliminary data inspection" purposes. Observe the rapid growth and subsequent decay of energy in the peak just below 0.1 Hz during the 2200-0030 time span (which has a different ordinate scale than the other two subfigures) and the emergence of the higher frequency secondary peak during the last three hour interval.

These spectra were based on 256-point FFT lengths, which corresponds to 8.5 minute subrecord lengths. It is instructive to explore how useful these spectral estimates actually are. Several aspects must be examined to answer this. The peak period in Figure 7.10 was 15 seconds; thus, each

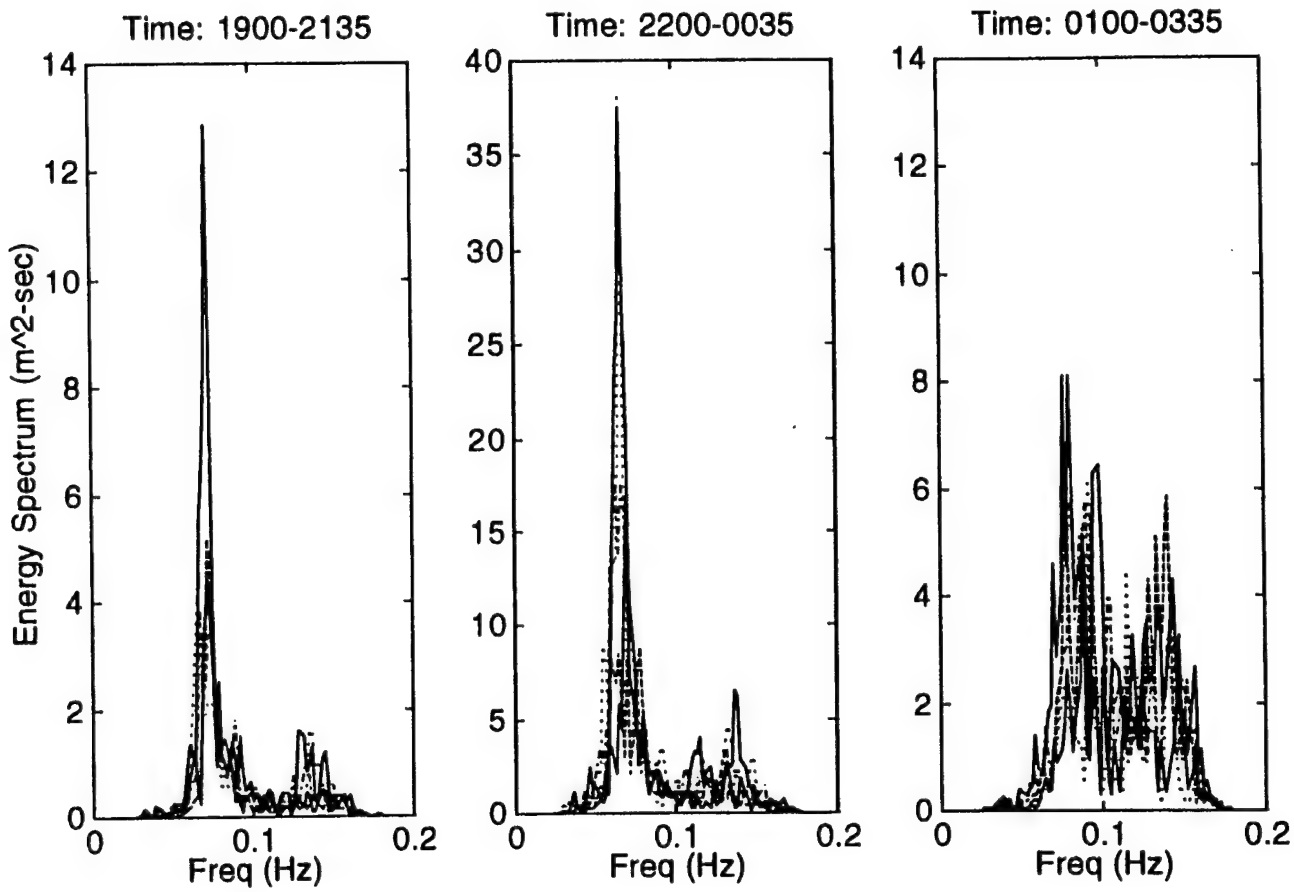

Figure 7.10 Representative Spectra over Half-Hour Intervals for Hurricane Bob

segment contained approximately 34 wave cycles. This is the minimum number of cycles recommended per FFT to produce reliable estimates. However, there were only 4 ensembles in each half hour spectral average which is far short of the 20 to 30 degrees of freedom recommended for standard spectral estimates (e.g., Goda (1985)). If this FFT segment length was retained to quantify this wavefield, then further algebraic frequency averaging would be required to achieve the minimum 20 degrees of freedom (in this case, averaging over every 5 frequency bins and reducing the frequency resolution accordingly). Assuming that the frequency resolution after such averaging was not considered acceptable, then the only remaining parameter to change is the FFT segment length to increase the number of ensembles. Simple inspection of Figure 7.10 clearly shows that the spectrum is changing on a time scale even shorter than this reference half hour segment length, so the nonstationarity of the signal makes increasing the number of averaged ensembles an unacceptable option. This is the classic "time-frequency ambiguity" and in this case it leads to the conclusion that spectral tools are applicable only as preliminary analysis tools to nonstationary storm events such as Hurricane Bob. Fortunately, HPT estimates will be shown to be quite informative for this event.

7.4.2 HPT Parametric Studies

This subsection has two objectives for this highly nonstationary wave signal: (1) demonstrate that HPT results are robust with respect to the

choice of segment length, and (2) introduce how HPT is capable of estimating the wavelength and incident wave direction. Data from the 1900-2150 wave data block (during the initial build-up of energy) were used as representative of the entire storm. The subsequent subsection will then investigate how HPT can be used to quantify the wave field corresponding to this storm.

The only independent parameter for HPT analyses is the segment length, and the objective here is to establish the optimum segment length. As with many other numerical techniques, this is a compromise since long segment lengths increase frequency resolution but short lengths improve stationarity and computer time. The most direct approach is to analyze the data with different lengths and visually inspect the results. Figures 7.11a, b, c, and d show various combinations of results for Gage 111 corresponding to the HPT analysis values in the following Table:

Descriptor	Segment Length (min)	Segment Shift (min)
Reference	10.0	2
Longer	12.5	2.5
Shorter	8.0	1.6
Doubled	20.0	4.2

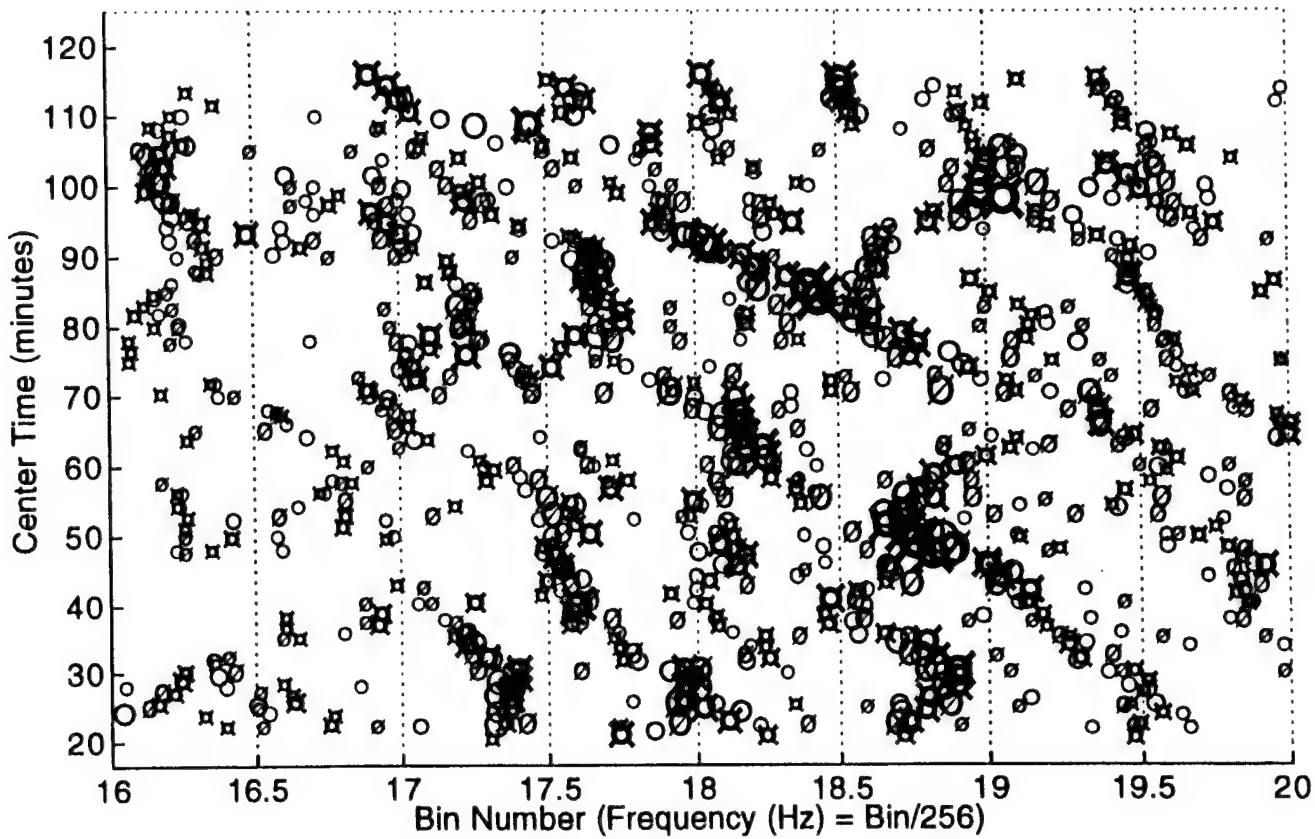

Figure 7.11a HPT Component Evolution for 3 Different Analysis Lengths: circle = 10; circle with slash = 12.5, and circle with x = 8 minutes, respectively.

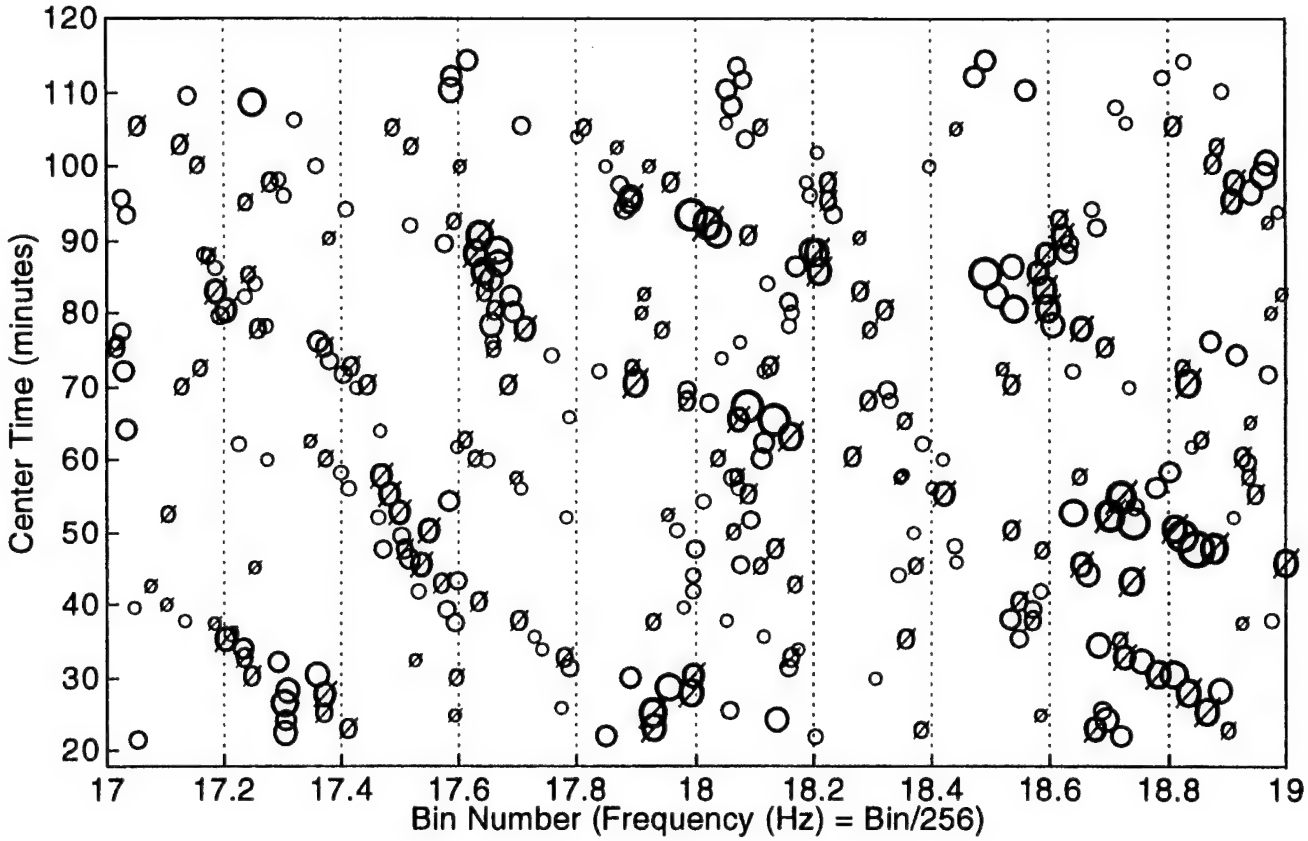

Figure 7.11b Detail of HPT Component Evolution for 2 Different Analysis Lengths: circle = 10; circle with slash = 12.5 minutes, respectively.

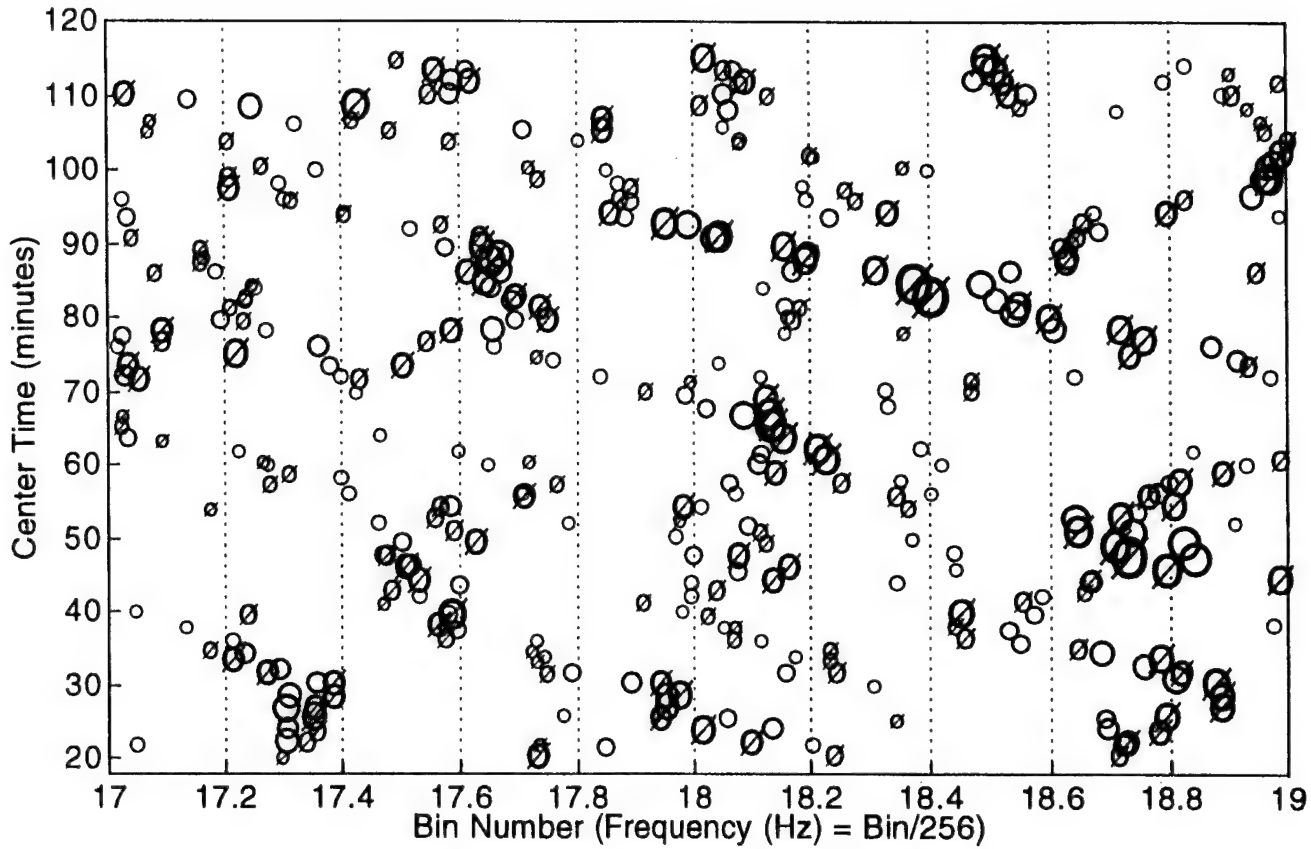

Figure 7.11c Detail of HPT Component Evolution for 2 Different Analysis Lengths: circle = 10; circle with x = 8 minutes, respectively.

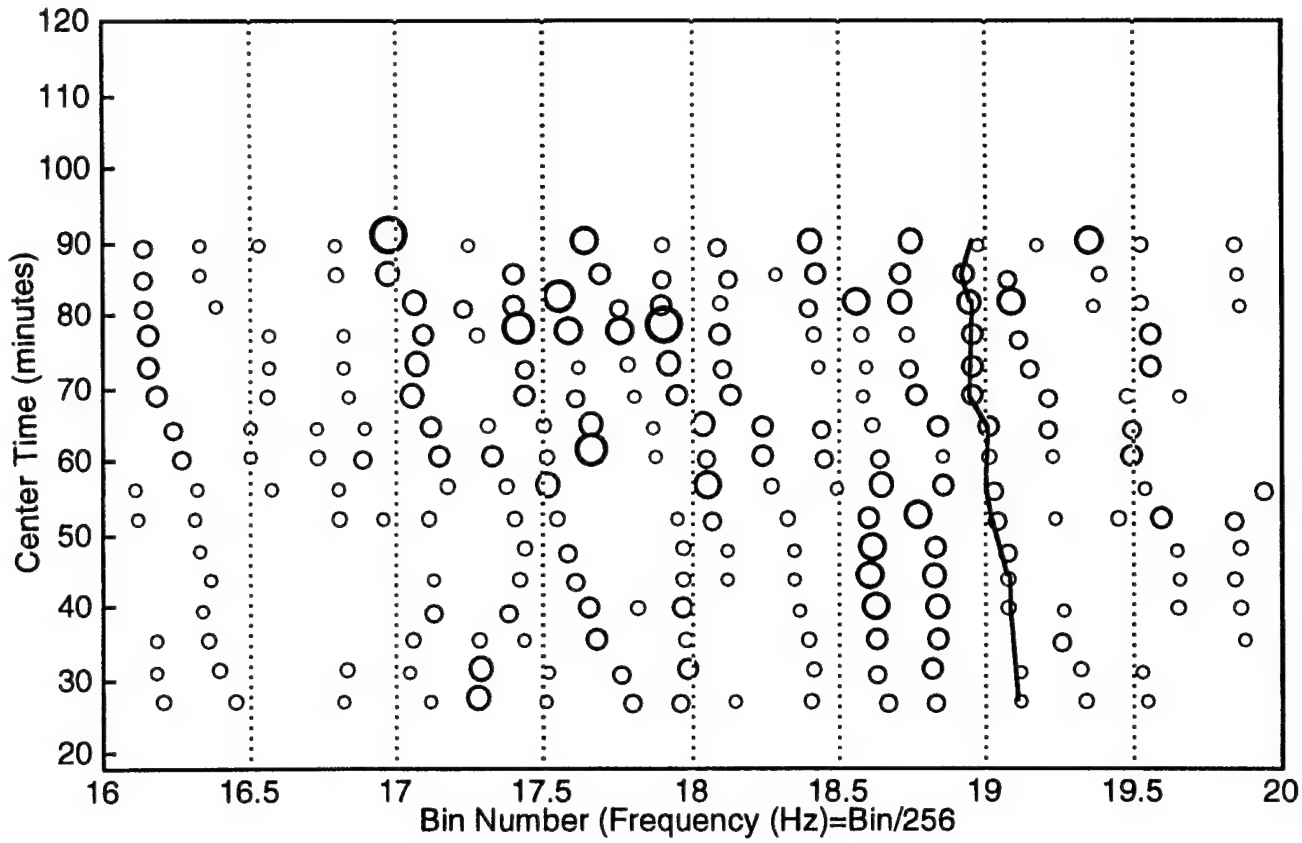

Figure 7.11d HPT Component Evolution for Doubled Analysis Length

Figure 7.11a shows that all of the results for the 8, 10, and 12.5 minute segments are reasonably consistent; packets are generally superimposed, and there are many areas where none of the techniques found any appreciable components. Figure 7.11b shows that the reference and longer segment results are indeed similar, while Figure 7.11c shows the shorter segment results starting to diverge (for example, observe the estimated amplitudes at position (18.4, 85)).

Figure 7.11d shows the wavefield evolution corresponding to the doubled segment length. It is apparent that the wave packets exhibit much less variance compared to the previous results. So while it is possible to conclude that these doubled results are inconsistent with the previous shorter-length results, it is still unclear as to which segment length best represents the physical wavefield.

As before, the most definitive test for the optimum segment length comes from inspection of phase continuity for representative packets. Figure 7.12 illustrates HPT packet evolution across the spectral peak region using the reference (10 minute) results. Figures 7.13a and b illustrate the evolution of one representative packet with significant energy indicated by the solid line in Figure 7.12. This packet shows reasonably constant frequency, with an amplitude that grows then decays similar to a classic wave "beat". The phases are very consistent except for a short time between 72 and 80 minutes. Besides the fact that the

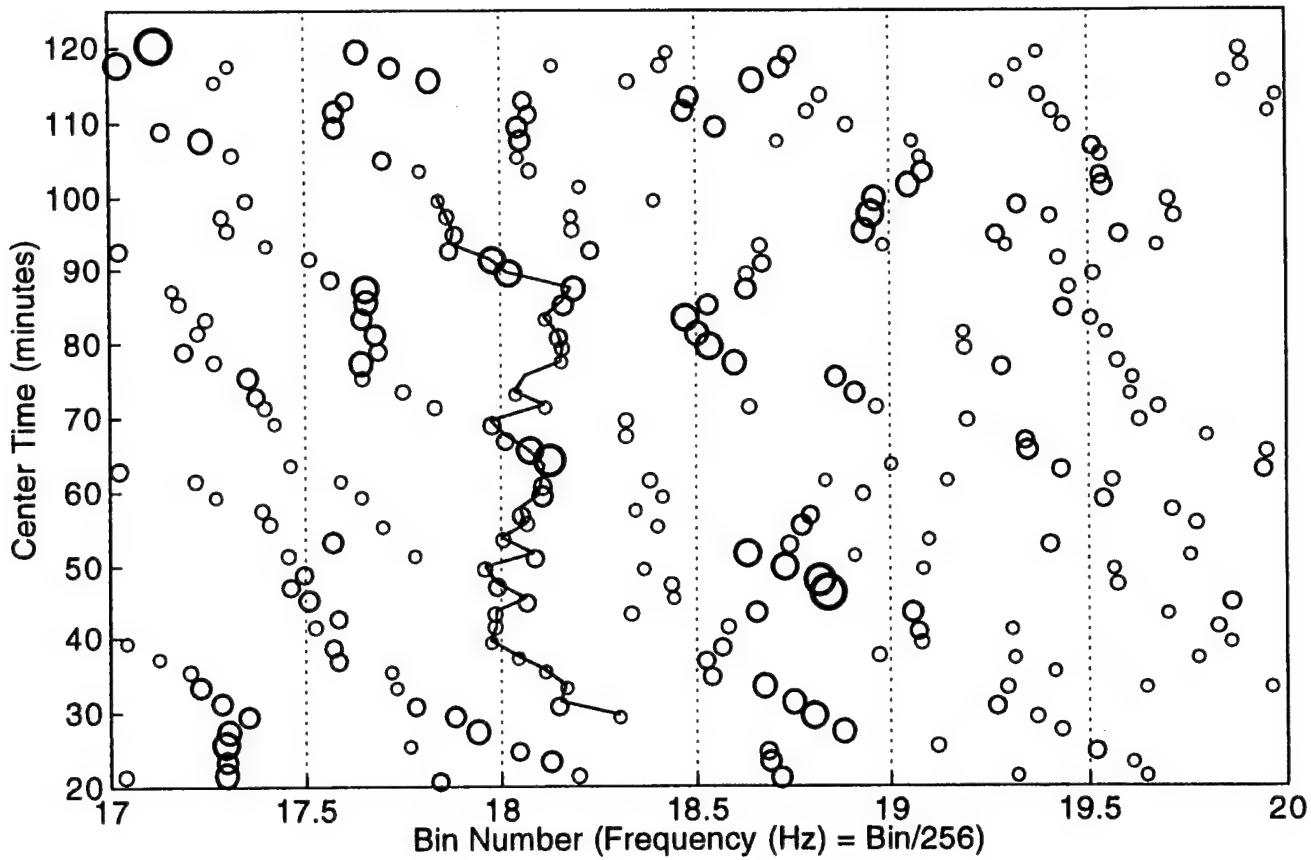

Figure 7.12 Wave Packet Evolution prior to Hurricane Bob

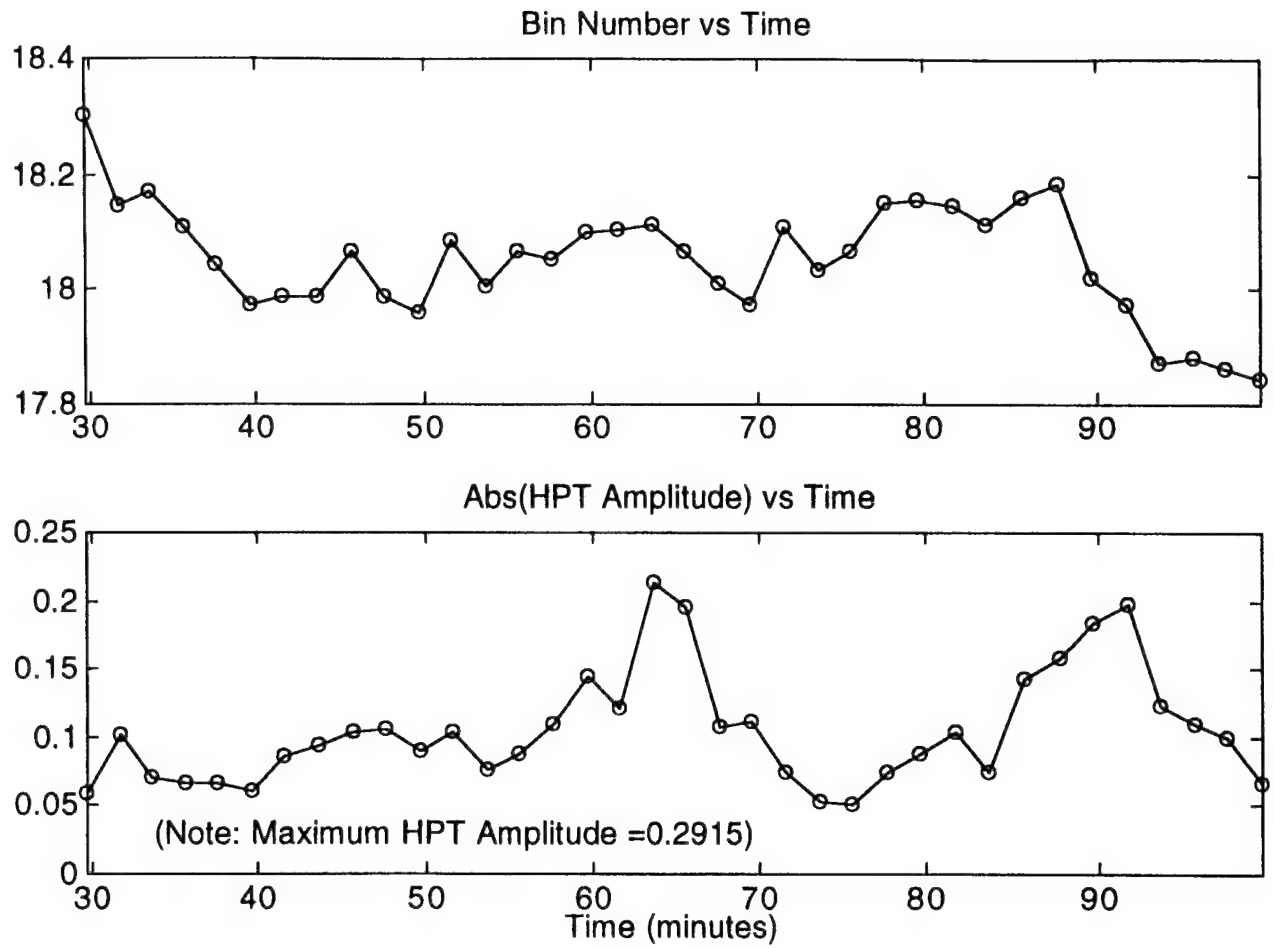

Figure 7.13a Representative Wave Packet Evolution prior to Hurricane Bob.

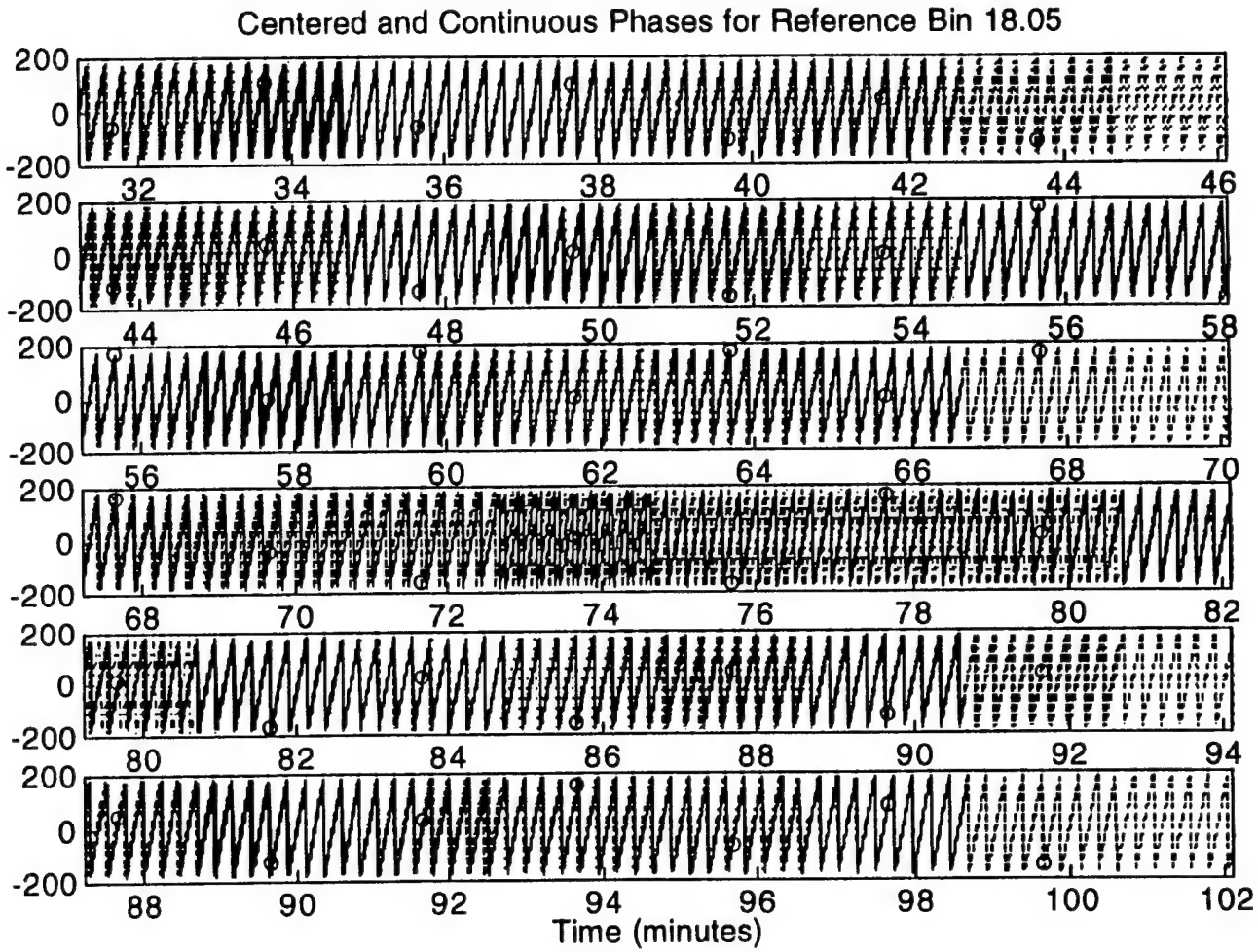

Figure 7.13b Phases of Representative Wave Packet prior to Hurricane Bob.

amplitudes are somewhat small and therefore could be unreliable, the reason for this poor agreement is unknown; it could be argued that a lower frequency packet was "created" around the same time, and that this new packet somehow effected the phases for the subject packet. Note that the total duration of this particular packet is over 70 minutes. Since the HPT segment length is nominally 10 minutes and at most 15 minutes accounting for the forward and backward time shift extensions, the phase continuity in Figure 7.13b cannot be attributed to numerical correlation effects, so it is proposed that it reliably denotes a physically present discrete wave packet.

The next step towards identifying the optimum segment length for HPT was to inspect phases for the packets estimated by the doubled HPT segment analyses shown in Figure 7.11d. A representative phase continuity function (for the packet that is centered on Bin 19) is shown in Figure 7.14. It is quite apparent that these estimates do not represent a single sinusoidal component, especially so considering the large 80 percent overlap between adjacent HPT analyses. Based on these phase results, the 10 minute segment length was chosen for the wavefield studies.

By way of contrast, Figure 7.15 compares HPT (10 minute) and FFT results along with the wave record for Gage 111. The FFT results, based on 512 point = 17 minute transforms, are far less informative and qualitatively quite different compared to the HPT results.

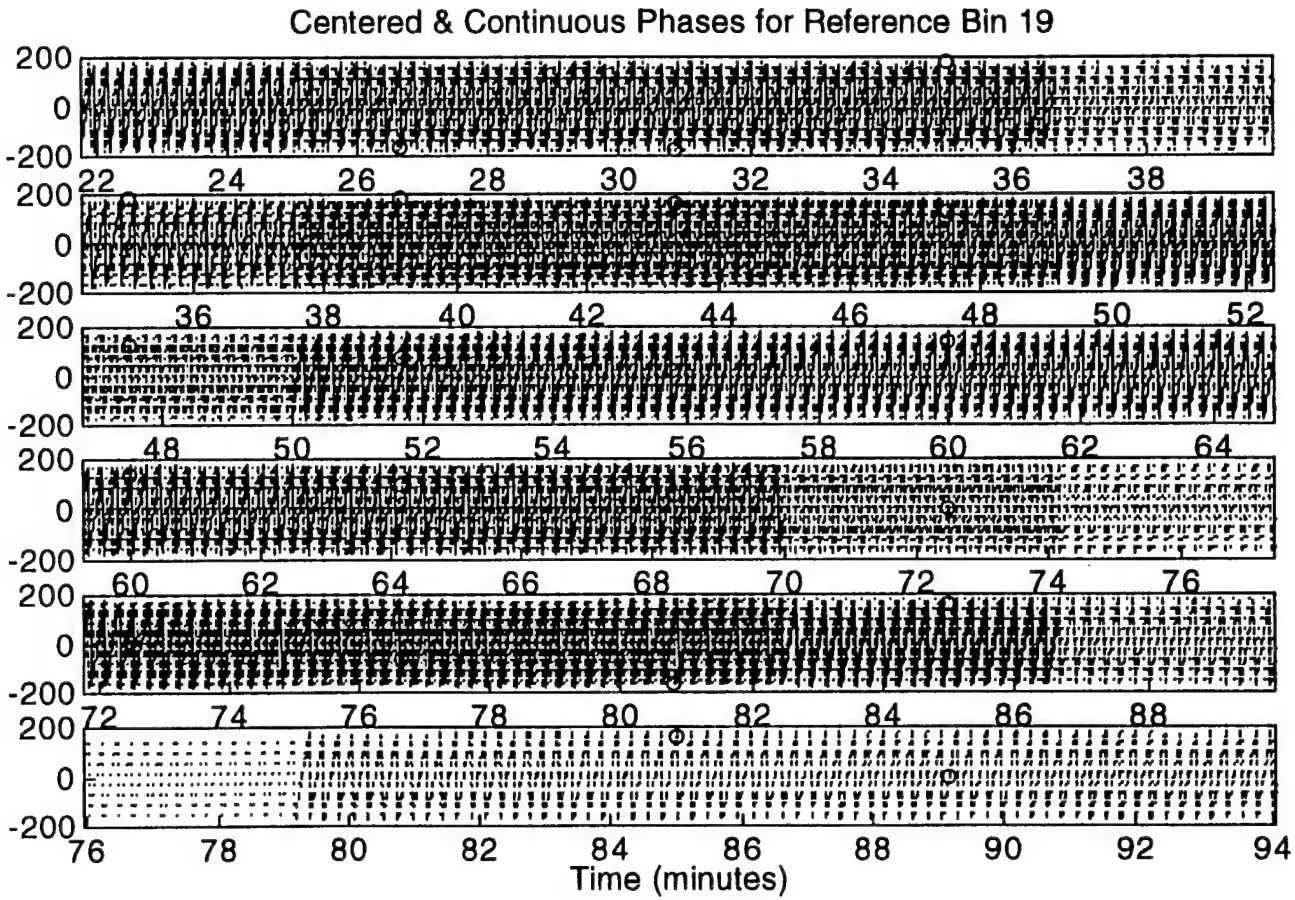

Figure 7.14 Phases of Representative Wave Packets for Doubled Segment Analyses prior to Hurricane Bob.

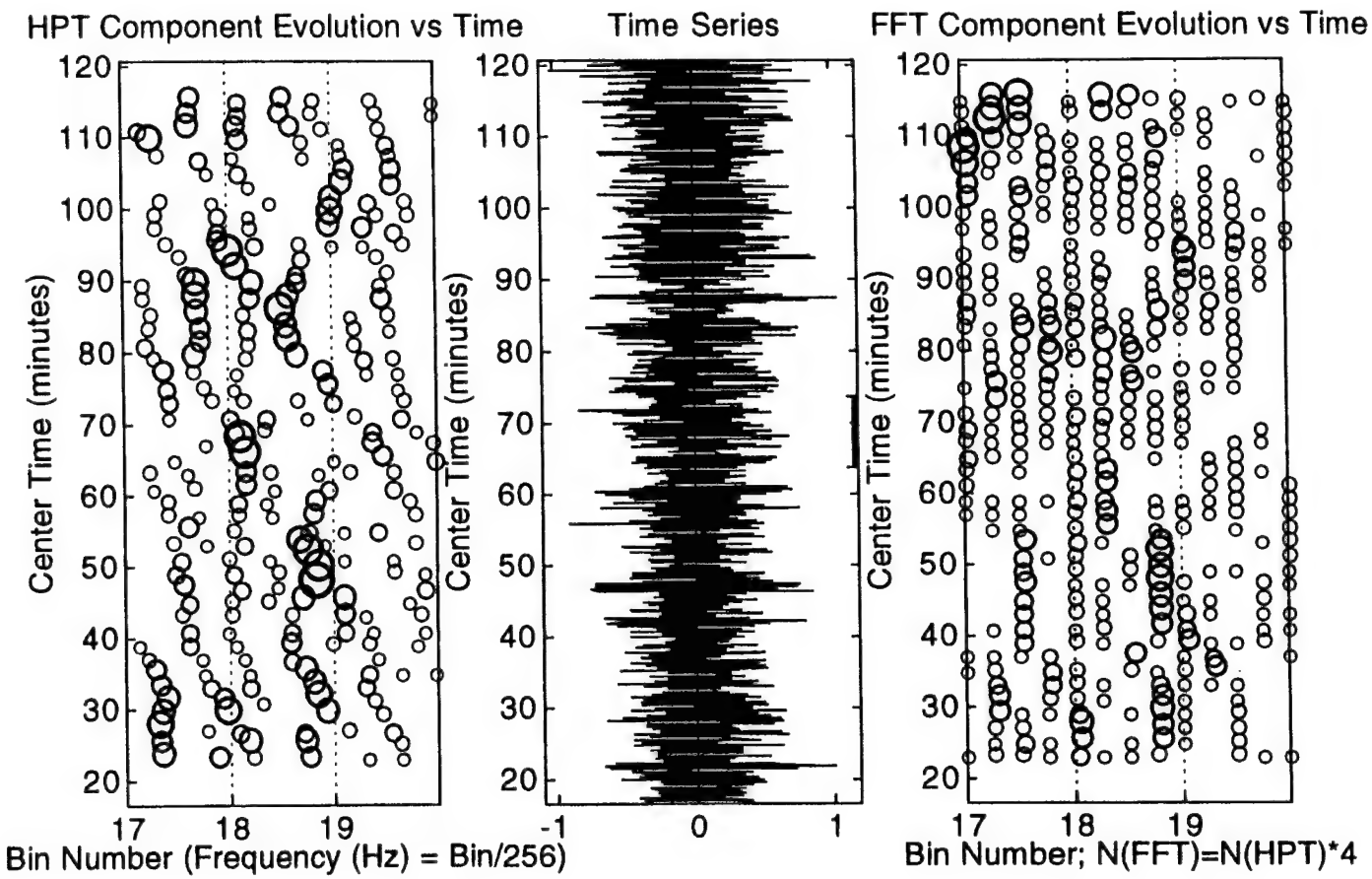

Figure 7.15 Comparison of HPT and FFT Component Evolutions during the initial stages of Hurricane Bob.

The third check for evaluating the HPT results comes from relative comparisons among neighboring gages, which complements the temporal checks with spatial checks. Recall that all wave records show correlations that are a function of the separation, less so orthogonal to the direction of wave advance (see Chapter 2 for the discussion on short crestedness and the coherence of the wave field). For example, Figure 7.16 has correlation functions for gages separated in orthogonal directions by 130m; since the incident direction was within 30 degrees of the East-West array axis, the two axes are approximately aligned in-line (upper figure) and orthogonal (lower figure) to the direction of wave advance. Observe that the correlation functions appear as damped sinusoids, consistent with the narrowbanded nature of the signal, and that the maximum correlation orthogonal to the wave advance is significantly lower compared to the in-line maximum.

Figure 7.17 presents the HPT estimates for the same gages used in this correlation study (approximately 0415-0425, August 18, 1991). Most of the frequencies are very consistent among the gages, with some scatter evident right at the peak frequency. Amplitudes show moderate variances.

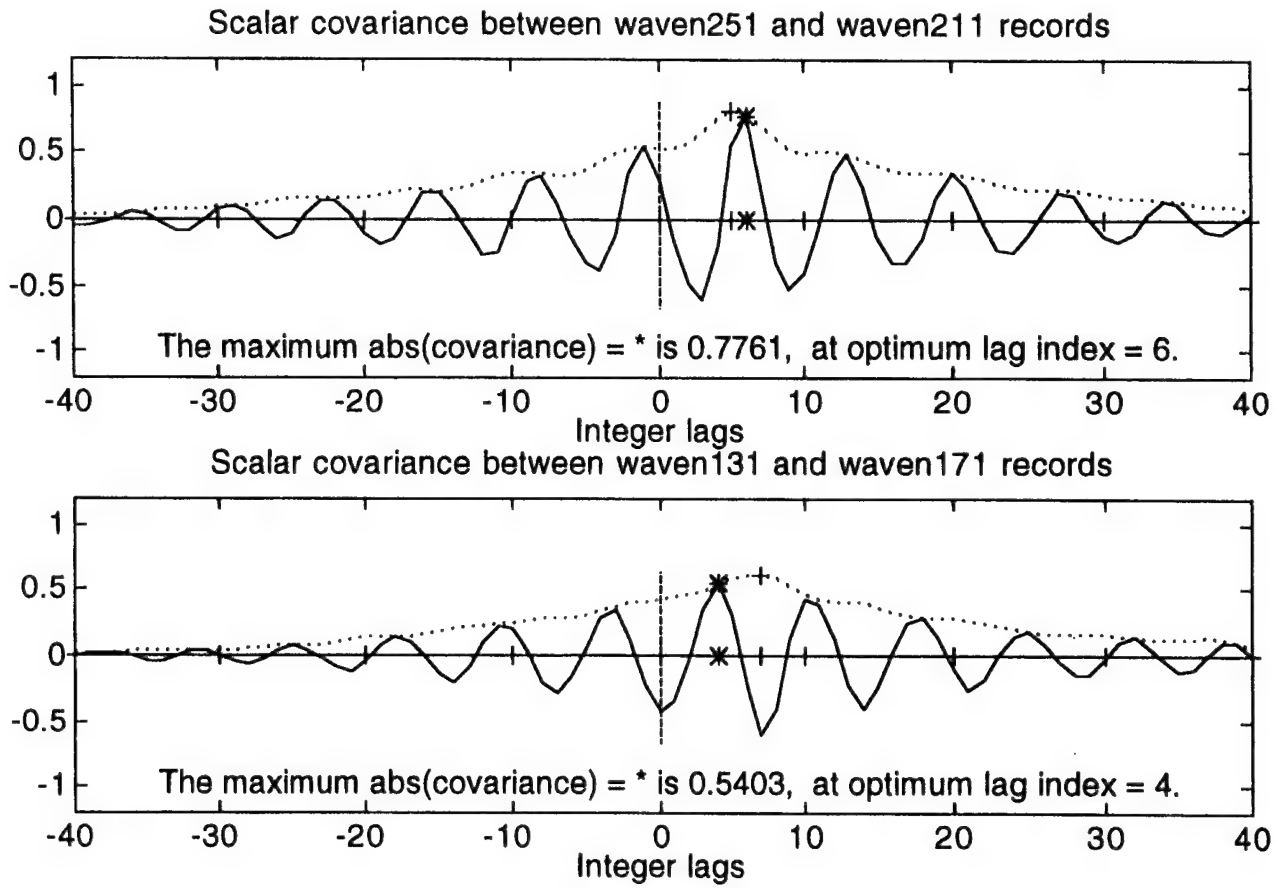

Figure 7.16 Representative In-line and Orthogonal Correlation Functions prior to Hurricane Bob.

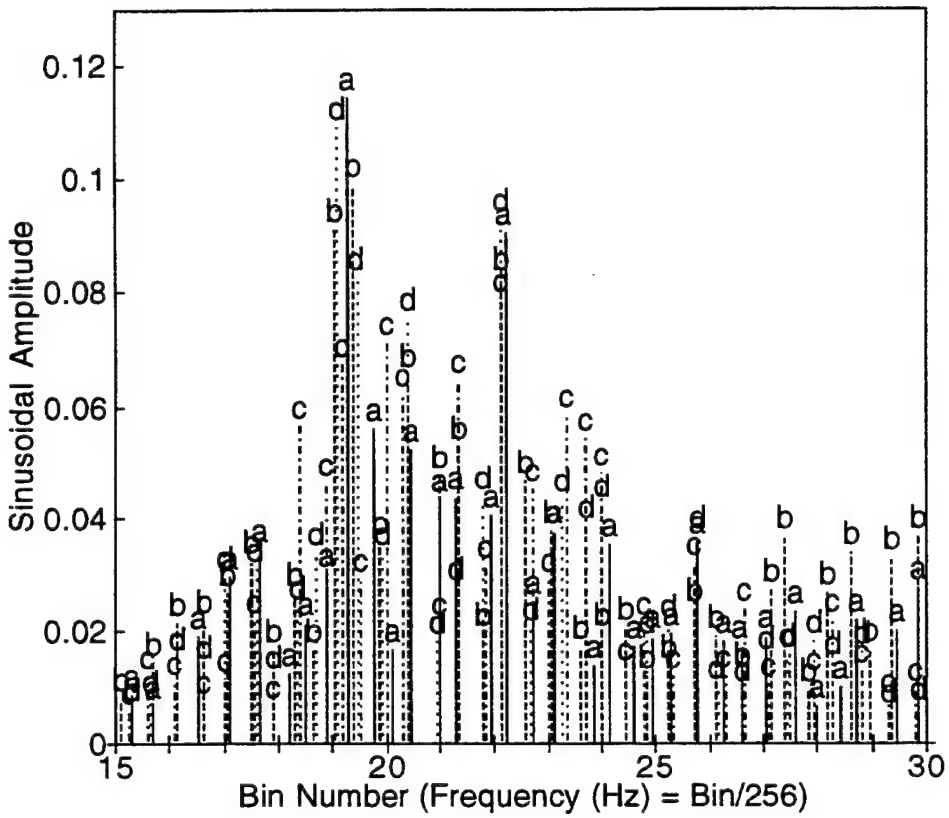

Figure 7.17 HPT Estimates for Gages at 130m spacing;

a = Gage 131, b = 251, c = 161, and d = 211;

The variability present in Figure 7.17 is most likely due to spatial non homogeneity in the wave field for some packets. HPT allows for further resolution of this variability by comparison of equivalent component (i.e., packet) evolutions between any gages of interest. Several examples are presented next. The first example shown in Figures 7.18a and b compares HPT estimates for Gages 251 and 211, which are predominantly in-line with a spacing of 120m. (Note: the symbol diameters

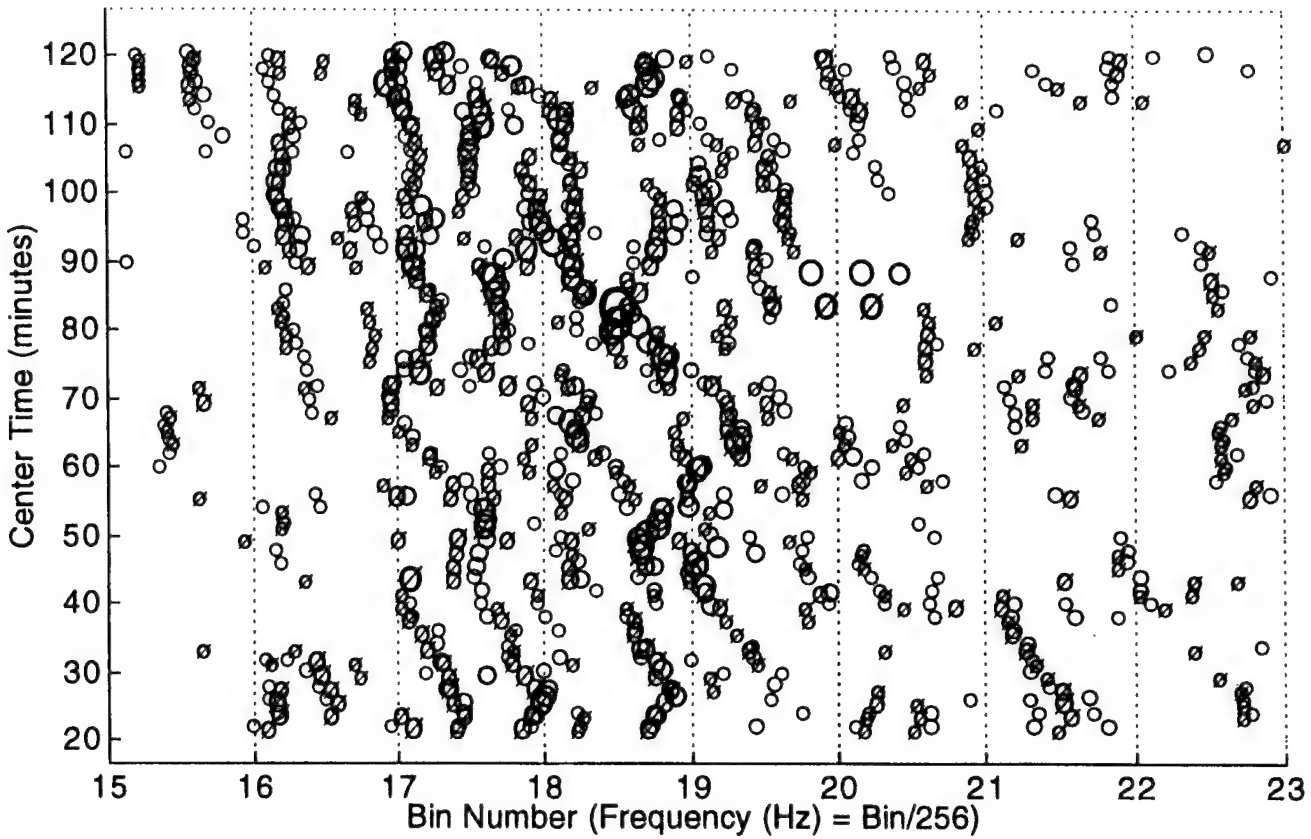

Figure 7.18a HPT Component Evolution over Main Bandwidth of Energy for Gages 251 (o) and 211 (ø) prior to Hurricane Bob.

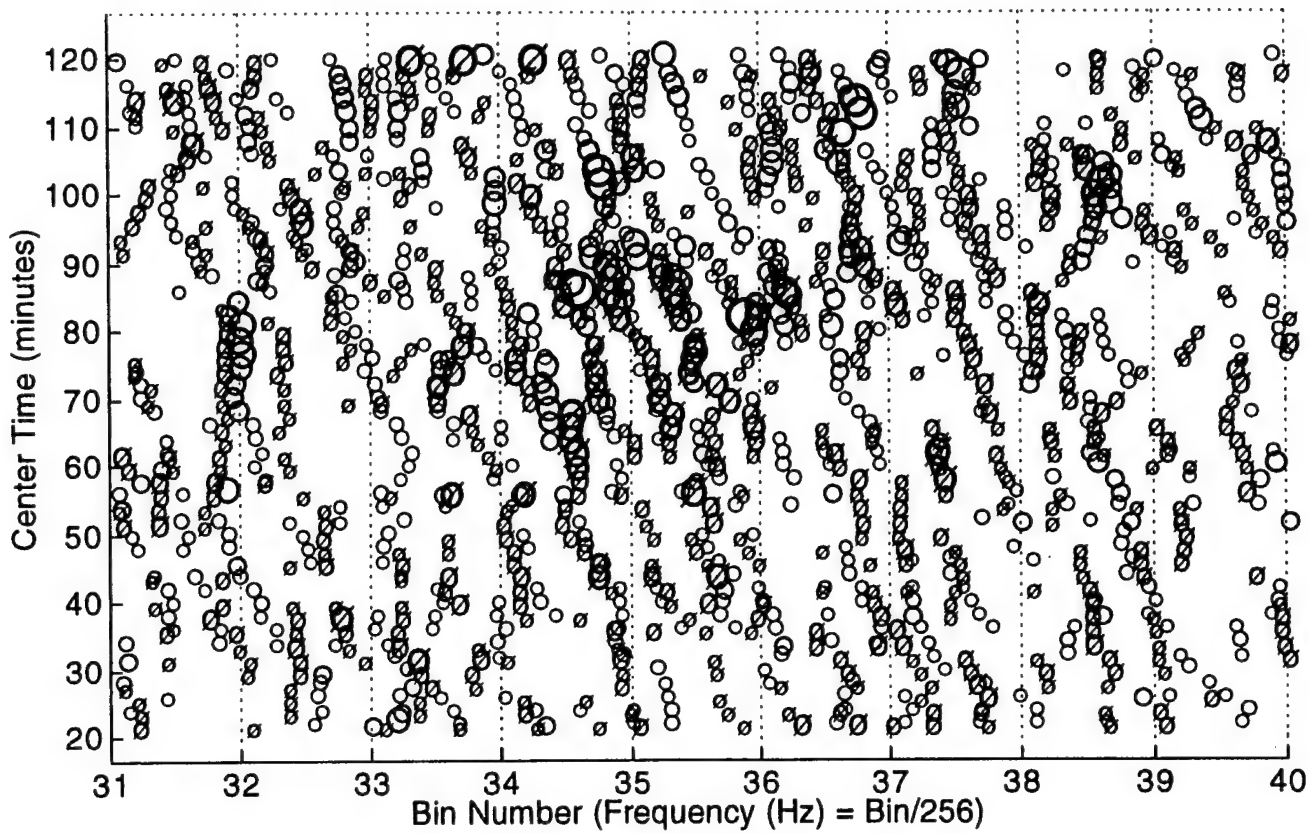

Figure 7.18b HPT Component Evolution over Secondary Bandwidth of Energy for Gages 251 (o) and 211 (ø) prior to Hurricane Bob.

corresponding to amplitude in each figure are normalized relative to the maximum amplitude over that customized frequency and time span; otherwise, evolution plots outside of the peak energy would show very few estimates with significant amplitude and would accordingly be uninformative. This explains why pairs of plots like Figure 7.18 show comparable amplitudes when the signal is known to be very narrowbanded.)

The frequencies, amplitudes, and duration of the wave packets estimated within the primary bandwidth of energy (Figure 7.18a) are very consistent between the two gages, while the packets over the secondary bandwidth are less consistent; Figure 7.19 details two equivalent packets from Figure 7.18a. The reduced consistency at higher frequencies in Figure 7.18b is expected since the gage separation is approximately one wavelength relative to the peak frequency but multiple wavelengths for the frequencies in the second figure. Or, it could be hypothesized that packets tend to self-organize over a relatively constant number of wave cycles, resulting in run lengths inversely proportional to frequency. This was not explored in this study.

The second example, shown in Figure 7.20, compares HPT estimates for Gages 111 and 191, which are predominantly orthogonal to the direction of wave advance with a spacing of 165m. This wider separation reduces the

maximum correlation coefficient to 0.48, which is reflected in the reduced consistency in the wave packets (although there is still a general

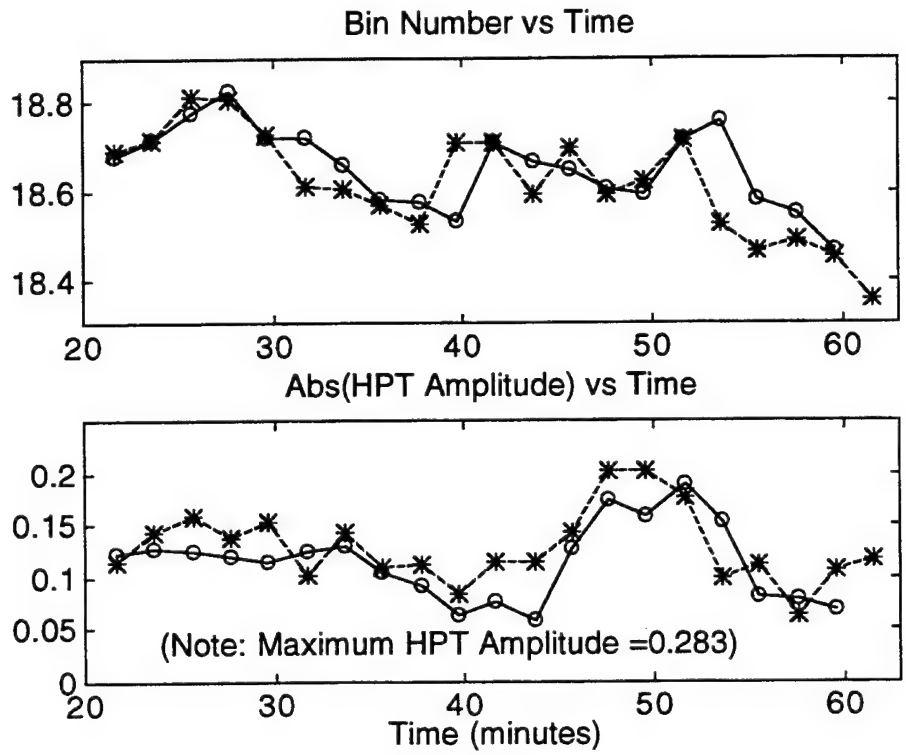

Figure 7.19 Comparison of Equivalent Wave Packets for Gage 251 (o) and Gage 211 (*)

correlation for many of the packets). Figure 7.21 shows equivalent packets for these two orthogonal gages, which appear to be reasonably consistent considering this gage separation.

The directionality of the waves was assumed for these previous examples based on FRF observations for the whole wavefield. However, just as with

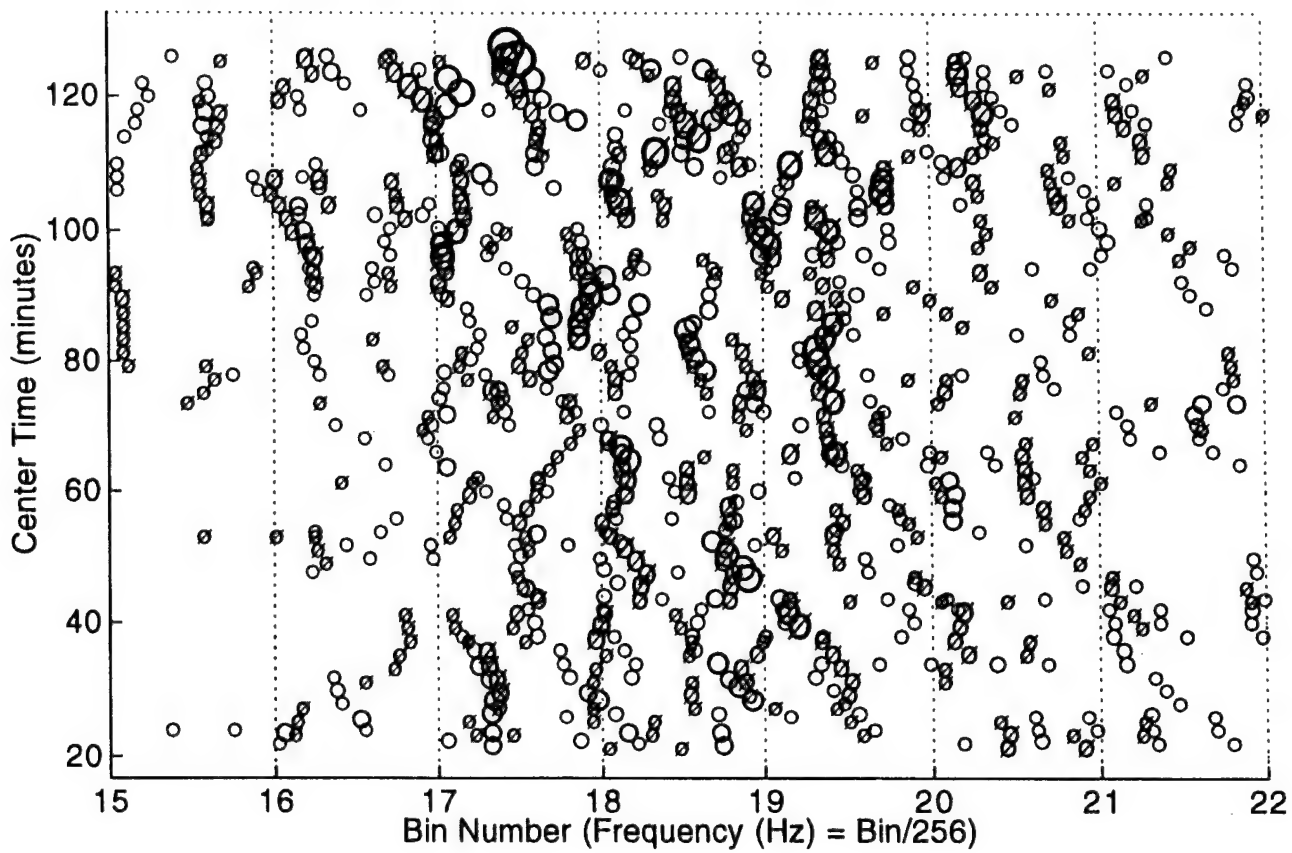

Figure 7.20 HPT Component Evolution over Main Bandwidth of Energy for Gages 131 (o) and 191 (ø) prior to Hurricane Bob.

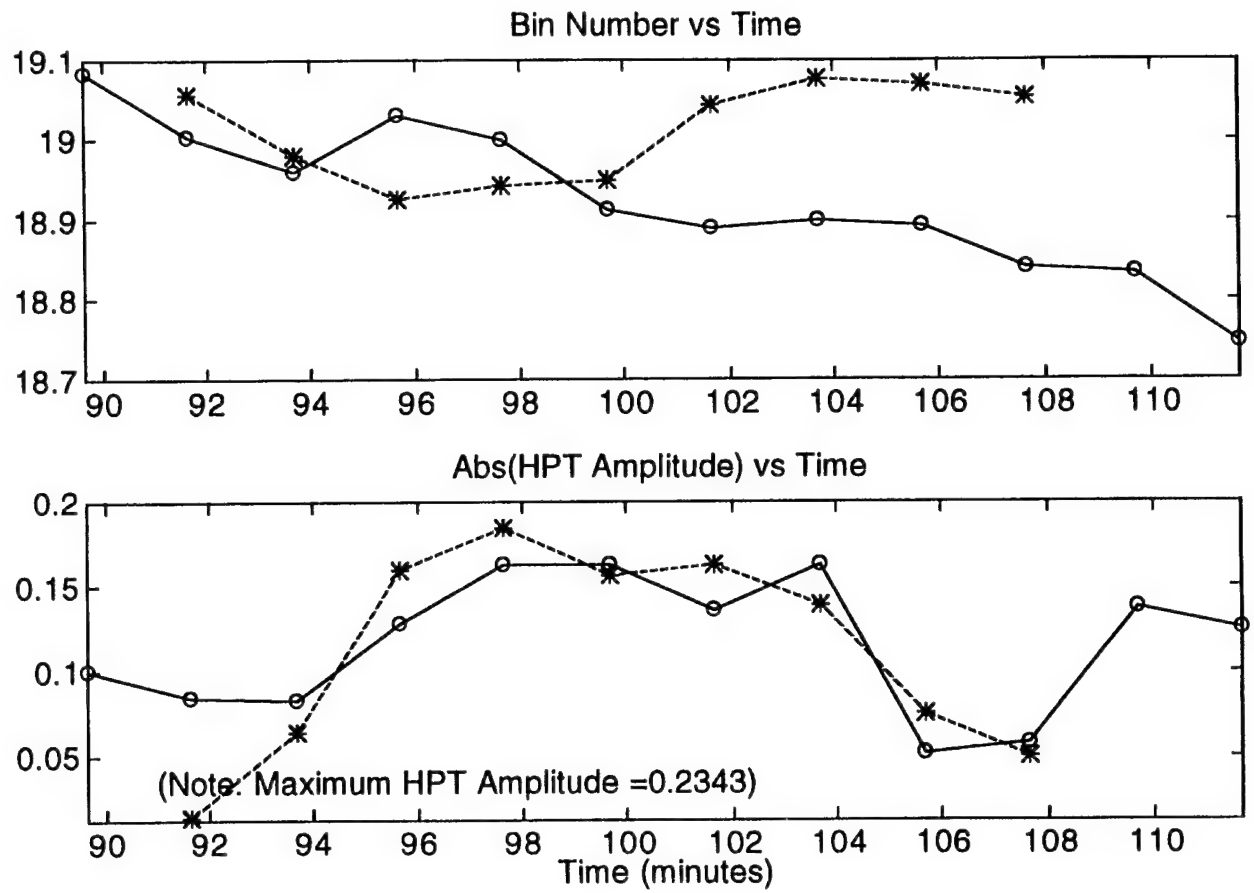

Figure 7.21 Representative Wave Packets for Gage 131 (o) and Gage 191 (*) prior to Hurricane Bob.

cross spectral estimation, the HPT frequencies, amplitudes and phases can also be used to provide quantitative estimates of the directionality versus frequency. But using HPT estimates for directionality is not as straightforward as it is using FFT-based estimates. First, it must be recognized that the spatial separations typically result in different frequency vectors for all of the gages as shown in Figures 7.17 through 7.21. These "raw" vectors must be compared and averaged to find a "best mean" vector which is not a trivial step. The criteria for selecting this mean vector from N gages is necessarily subjective. There are usually some well-defined frequencies where all N gages show good agreement, with clear gaps to the neighboring lower and higher frequencies; other frequencies may show a "dual" nature, with some gages estimating one frequency but others estimating two closely-spaced frequencies (typically encountered when the segment straddles the node of the packet envelope); while for some frequencies the component may not have been significant for a proportion of the gages so that a lesser number of gages are represented. Defining a "best, mean" vector requires choices for: (1) the frequency limits used to identify and collect similar frequencies, (2) the minimum number of gages necessary to define a frequency deemed common to all of the gages, and (3) the scheme for combining raw frequencies within a band (for this study, amplitude weighting was found to be most effective). With this new best vector so defined, it is used to define a total least squares basis matrix, and new amplitudes and phases are recalculated for all of the gages.

The estimated wavelength and the mean incident wave direction are found at each frequency by first plotting the HPT phase versus coordinate for each orthogonal direction. In most cases an adjustment step is required to unwrap these raw HPT phases (another necessarily subjective and non-trivial task). Note that the (unwrapped) phase versus coordinate is necessarily linear if the frequency and the phase are constant, which corresponds to a long crested, coherent wave packet. For these studies, least squares was used to fit a two-dimensional plane of phase versus gage coordinates. The ratio of the slopes in the two orthogonal directions then defines the estimated incident direction. The component wavelengths in both of the orthogonal directions are found at each frequency by using the component slopes to estimate the coordinate span corresponding to 360 degrees (one wave cycle), then combining these "projected" wavelengths to get the effective wavelength relative to the incident direction. As an aside, visual inspection of the phase continuity in the two orthogonal directions defines another technique for assessing the continuity of a wave packet.

The next set of figures and table illustrate this process. Figure 7.22 illustrates the "raw" gage frequencies (and amplitudes) for all 12 gages in the FRF array, near the vicinity of the peak energy, and starting at 1915. The "best, mean" frequencies are indicated by the circles on the x-axis.

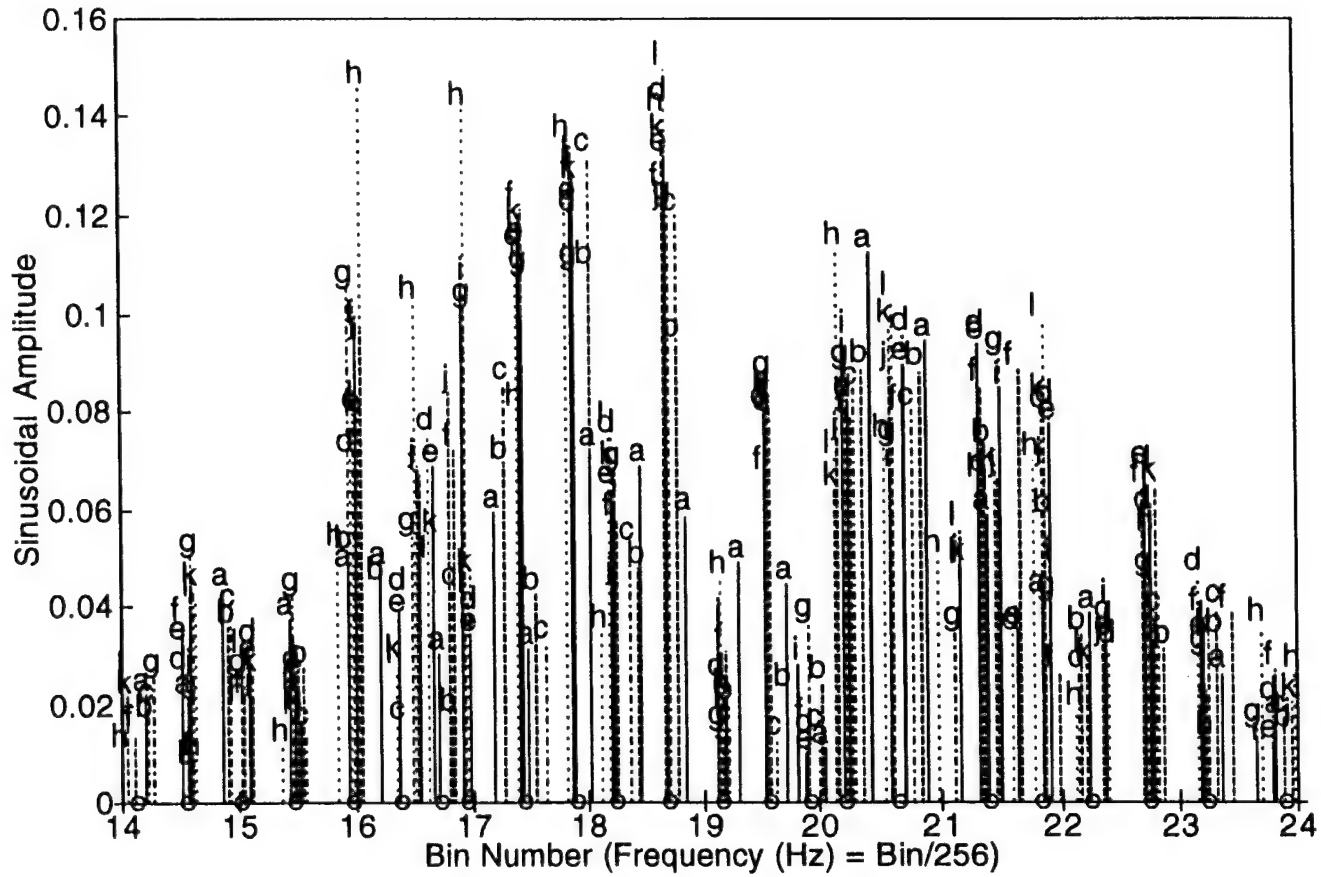

Figure 7.22 Illustration of Raw Gage Frequency Vectors (letters) and Best, Mean Frequency Vector (o symbols along abscissa).

Table 7.3 illustrates the quantitative information calculated at each frequency. The phase columns in the two subtables are defined as:

- "Raw" refers to the original HPT wrapped estimate.
- "Adjst" refers to the adjusted (unwrapped) equivalent; in this example the program shifted raw values by up to two cycles to produce an "optimum" fit relative to each orthogonal direction;
- "Fitted" refers to the best least squares fit based on a two dimensional plane fitted to all of the estimates, and
- "Error" refers to the deviation between the adjusted and best-fit phases.

Finally, error bounds appropriate for the linear fits are determined based on coefficients of determination used in linear regression for each orthogonal (north-south and east-west) direction, $R_N{}^2$ and $R_E{}^2$, respectively (e.g., Montgomery and Peck, 1992). Observe that the estimated and analytical wavelengths are essentially equal for this example case.

Figure 7.23 graphically displays the phase information from Table 7.3. The phase fit is seen to be very good in both orthogonal directions, implying that these HPT estimates do represent a coherent and long crested packet traveling through the wave field. This is significant, because the 52m wavelength corresponding to this example translates to relatively large

Gage	North-South Phase (degrees)		Best Linear Fit (degrees)	
	Raw	Adjst	Fitted	Error
191	4	-716	-702	15
181	106	-614	-603	10
171	179	-541	-533	7
111	156	-204	-232	-28
121	-173	-173	-205	-32
131	-141	-141	-161	-19
151	-56	-56	-49	7
161	-37	-37	22	59

Gage	East-West Phase (degrees)		Best Linear Fit (degrees)	
	Raw	Adjst	Fitted	Error
211	47	-673	-668	6
231	158	-202	-223	-21
131	-141	-141	-161	-19
241	-12	-12	-33	-21
251	77	77	94	17

Coefficients of Determination: North-South=0.988 & East-West=0.995

Projected Wavelengths (m): North-South= 127 & East-West= 57

Estimated Wavelength (m) = 51.8; Analytical Wavelength (m) = 52

Estimated Incident Direction (deg) = 114

Table 7.3 Example Directionality Estimate For Frequency = 0.148 Hz

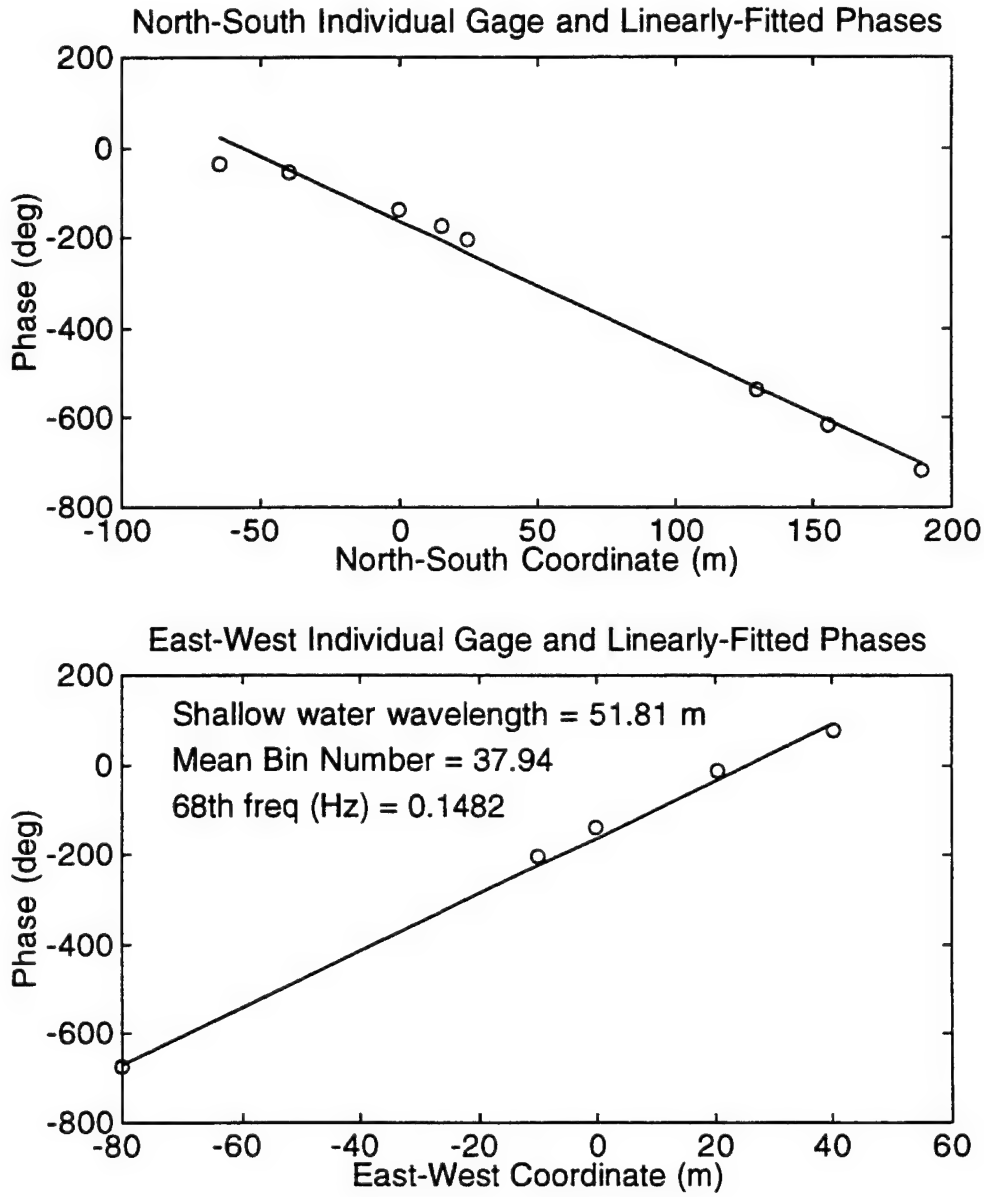

Figure 7.23 Example Adjusted and Fitted Phases for Frequency = 0.148 Hz.

normalized array dimensions of 5 and 2.5 wavelengths for the two orthogonal directions.

Figure 7.24 presents the statistically valid HPT estimated wavelengths versus frequency, with 90 percent error bounds, for three different HPT segment lengths and one FFT segment. Statistically valid is defined in a number of ways to accomodate various situations; such as: (1) valid oblique incident direction (both $R_N{}^2$ and $R_E{}^2$ greater than 0.707, or $R_N{}^2$ plus $R_E{}^2$ greater than 1.3), or a valid predominant array alignment (only $R_N{}^2$ or $R_E{}^2$ significant). Specifically, the normalized wavelength error bound is defined as directly proportional to the standard error of the corresponding component slope (using linear regression theory). The lower subfigure shows FFT-based wavelength calculations for reference. Only one (raw) transform is used to be consistent with the single segment used in the HPT calculations. Inspection of these four subfigures shows that: (1) the HPT estimates are not sensitive to these choices of segment length, and (2) that the HPT estimates are consistent with traditional FFT-based estimates.

This agreement between the HPT and FFT results was not initially expected. HPT frequencies correspond to physically-present wave packets in which phase continuity is maximized over space and time for coherent, long crested packets; since the component HPT estimates at each gage accordingly have *low bias*, then wavelength and incident direction estimates should have low bias as well. On the other hand, the FFT

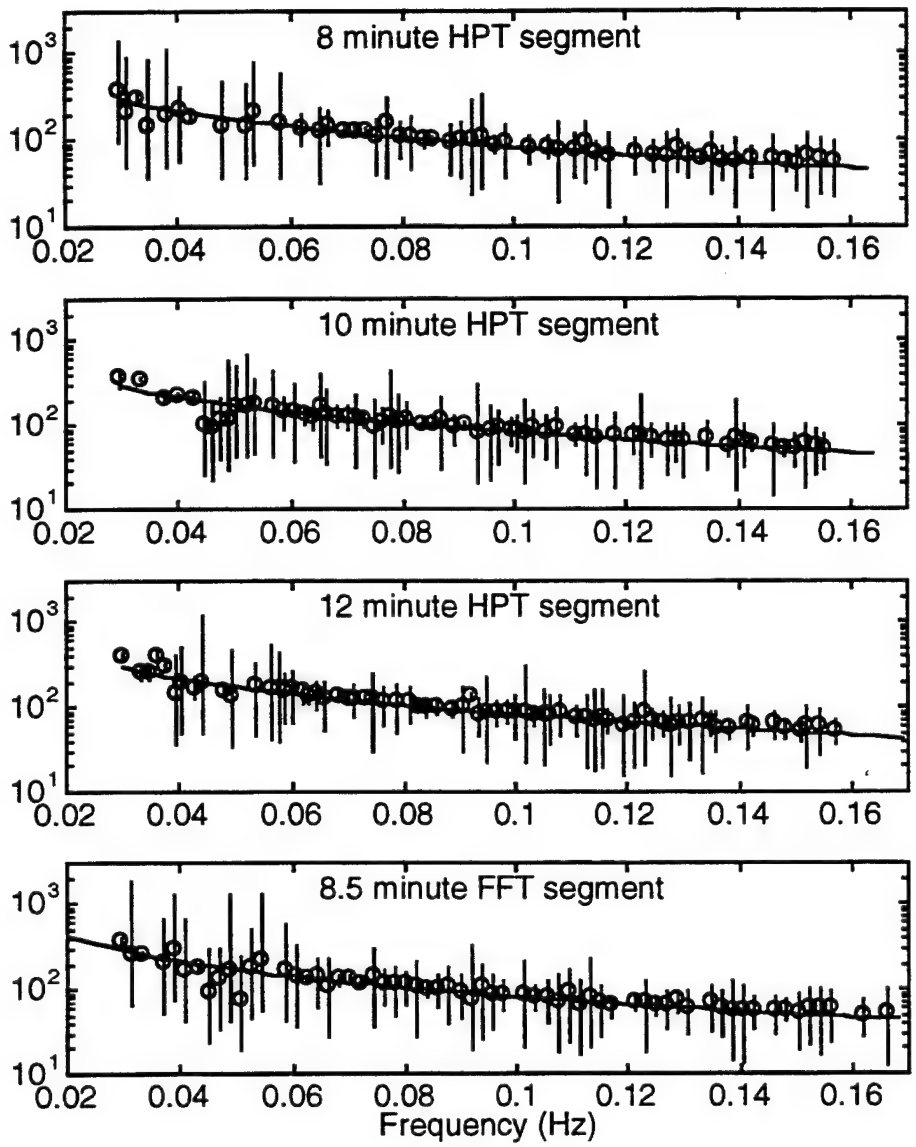

Figure 7.24 Estimated and Analytical Wavelengths versus HPT Segment Length, with FFT Reference

frequencies represent equally-spaced convenient (i.e., orthogonal) harmonics with no physical meaning, so the individual phase estimates among the gages typically do not correspond to any particular wave packets and hence are *generally highly biased*. The answer apparently is that, while the FFT estimates are all biased, they are all *consistently* biased, and hence all phase differences, and subsequent calculations for wavelength and direction, can still be unbiased. Thus, "two wrongs make a right" in the case of FFT-based calculations for wavelength (and incident direction on the next figure). If nothing else, this proves the power of orthogonal decompositions for solving real-world problems.

Figure 7.25 shows the HPT estimated incident directions for this example. Observe again that the variance and accuracy of the HPT estimates appear to be relatively constant for this range of segment lengths. The HPT estimates are also essentially equivalent to the FFT-based estimates. Both techniques detected waves with negative mean incident angles, which in this case correspond to reflected waves off the beach (observe the mirror image reflection angle).

It is therefore concluded from Figures 7.22 to 7.25 that: (1) the 10 minute segment length is valid for HPT studies, and (2) that in this case HPT does not provide new information not already available from the FFT for wavelength and incident direction estimates.

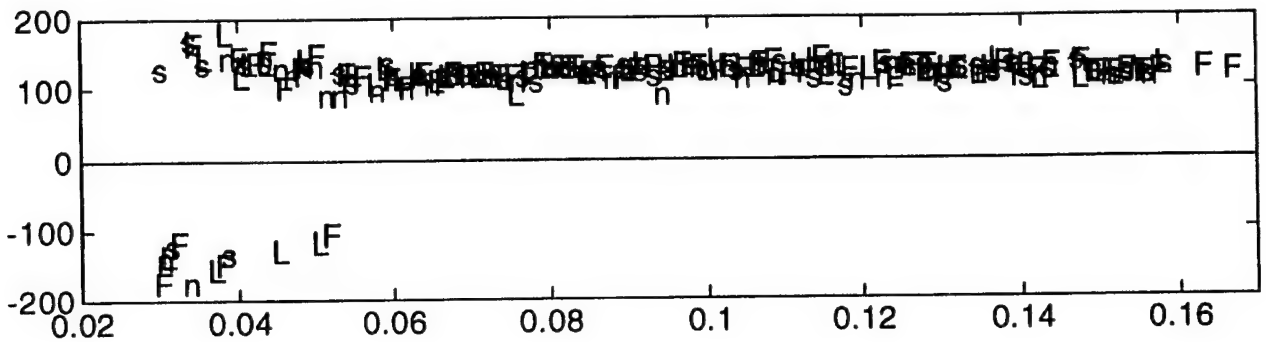

Figure 7.25a Estimated Incident Directions (Deg) versus Frequency (Hz) for FFT (F) and HPT analyses: L, n, s = 12, 10, and 8 minute segment lengths, respectively.

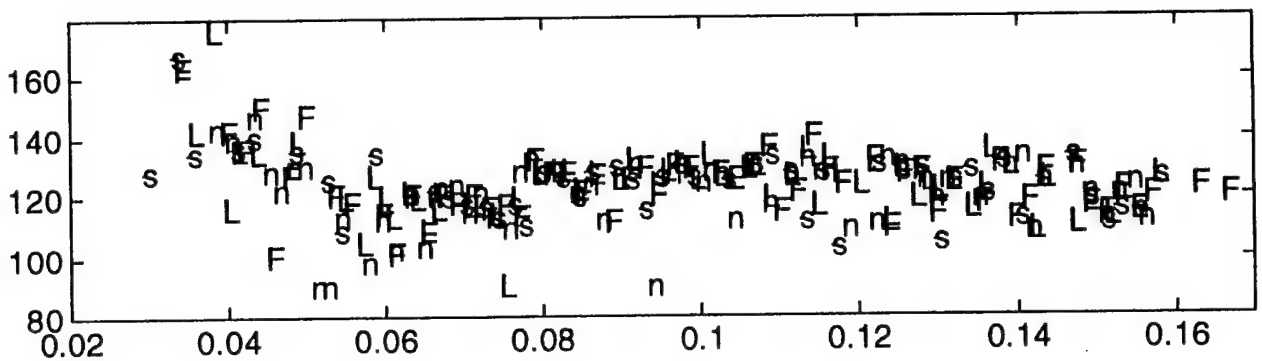

Figure 7.25b Detail from Figure 7.25a of Estimated Incident Directions.

These last few pages have examined whether HPT results are consistent and accurate (unbiased) measures of this growing wave field. Representative figures were presented to establish that HPT results are indeed consistent with expected packet behavior with translations in time and space.

The next subsection applies all of these HPT tools to a preliminary study of the wave field in Hurricane Bob.

7.4.3 HPT Investigation of Hurricane Bob Wavefield

As discussed in the first part of this section and shown in Figures 7.9 and 7.10, the wave field during Hurricane Bob had three distinct stages with differing stationarity as the storm progressed northward past the FRF site.

Figure 7.26 shows the HPT estimated mean incident wave directions at four times during the storm. The upper two figures are before the peak and show the waves coming from a southeasterly direction. Both lower figures are after the peak and show that the waves have shifted direction and are coming from the northeast, with a slightly increased variance. This information is necessary for decomposing the wave packet correlations into the two orthogonal directions relative to the wave advance.

Figures 7.19 and 7.22 and others illustrated packet evolution for the first block of wave data from 1900 to 2140, before the arrival of the main part of the storm, corresponding to the left subfigure in Figure 7.10. Figure 7.27 illustrates one frequency range of packet evolution for three gages just before and during the peak of the storm, corresponding to the middle subfigure in Figure 7.10. HPT frequency resolution was 0.27 bins, or equivalently, 0.0011 Hz. Both figures show a high degree of consistency among the gages; for reference, the maximum correlation coefficient is 0.87 between Gages 251 and 111, and 0.59 between Gages 251 and 191. Also,

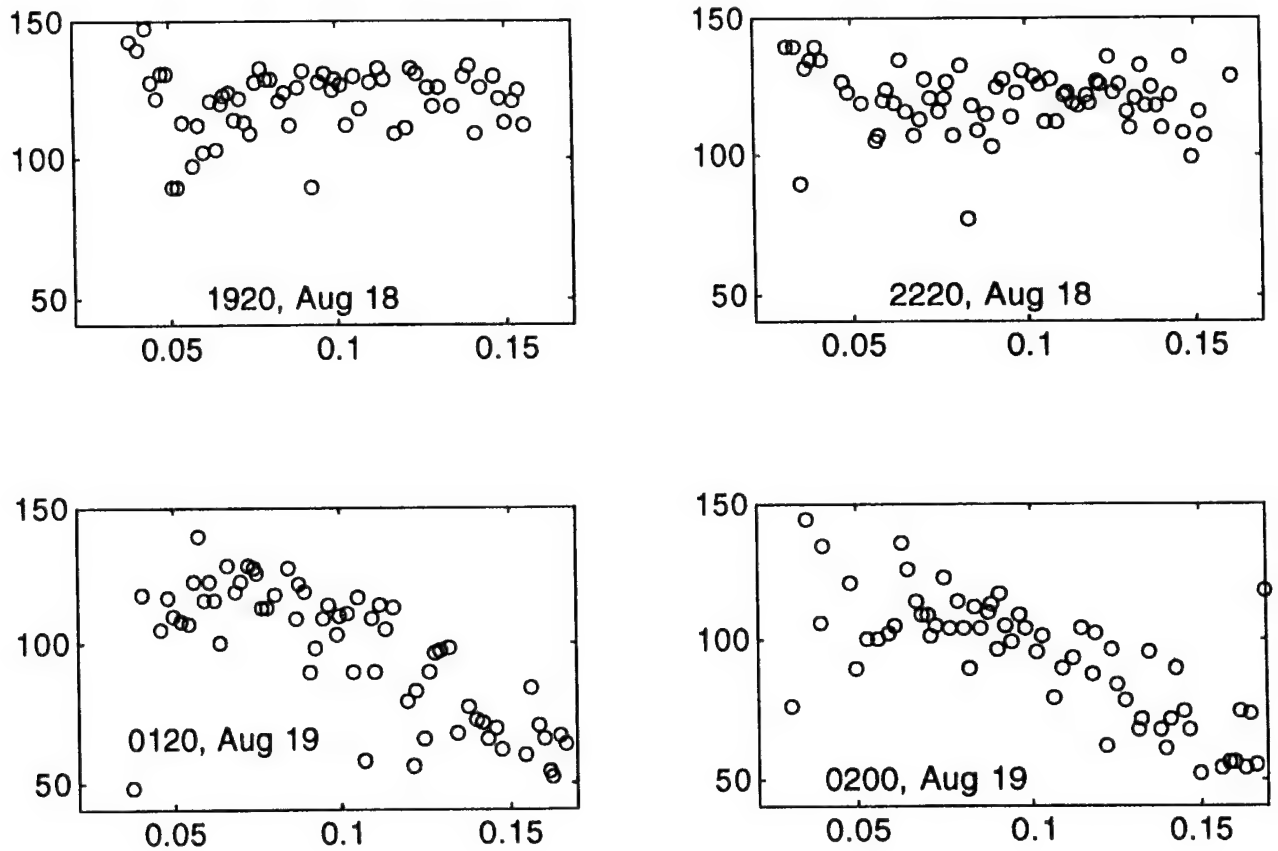

Figure 7.26 HPT-Estimated Incident Directions (Deg) versus Frequency (Hz) During Hurricane Bob.

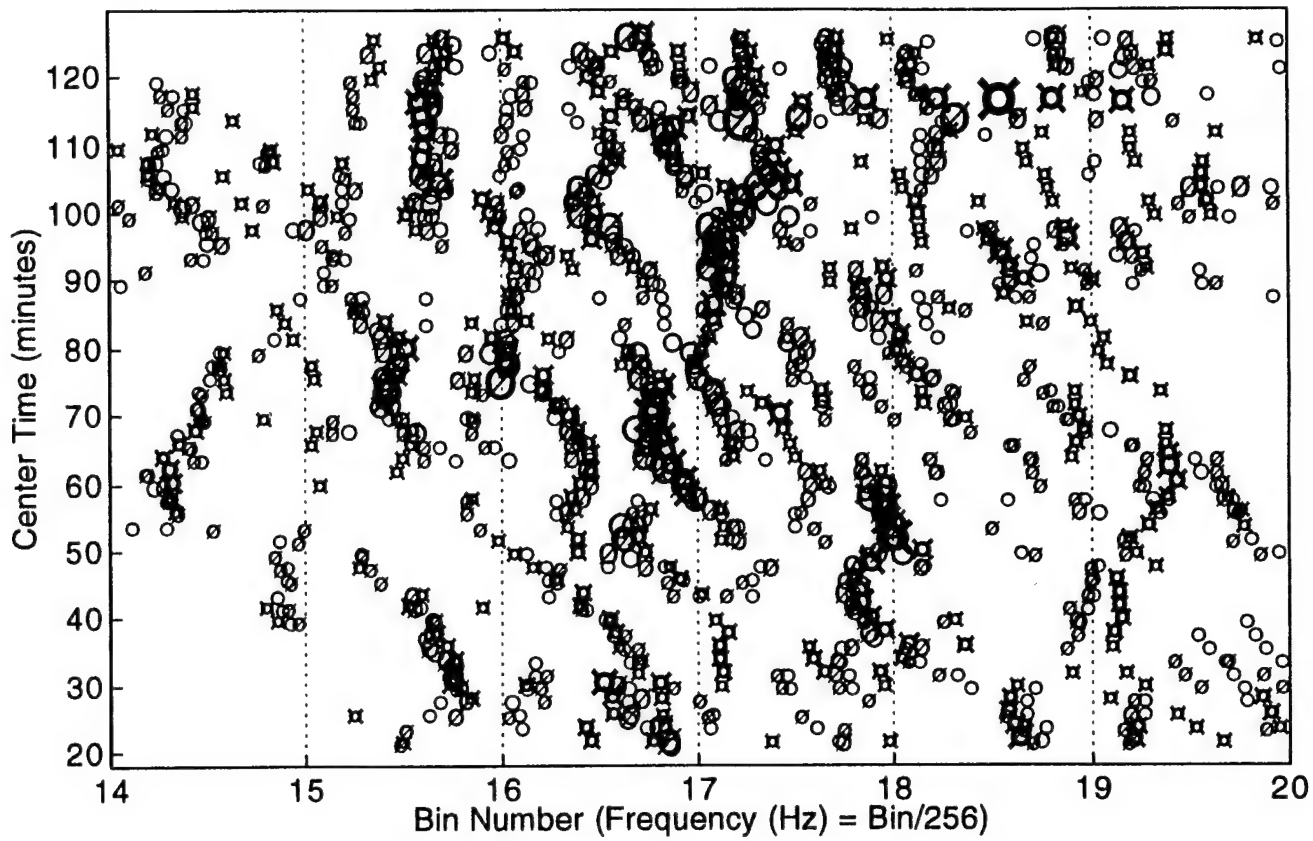

Figure 7.27 Wave Packet Evolution during the peak of Hurricane Bob for Gages 111 (o), 251 (ø), and 191 (o).

both figures show long durations for the packets, typically 20 to 30 minutes, with a high degree of overlapping.

Figure 7.28 examines one of the packets for all three gages. Further insight into the expected correlations among these packets can be found from the expected properties for this packet and the underlying wave. For a mean bin number of 15.7, pertinent parameters are calculated below using Dean and Dalrymple, 1984:

Mean Frequency ($\bar{f}$)	= 0.0613 Hz
Mean Period ($\bar{T}$)	= 16.3 seconds
Wavelengths:	
Deep Water (λ_0)	= 414.5 m
Shallow (λ)	= 77.3 m
$(kh) = (2\pi)depth/\lambda$	= 0.189
Deepwater Celerity $(C_0) = \lambda_0/\bar{T}$	= 25.7 m/sec
Shallow water Celerity $(C) = C_0 \tanh(kh)$	= 9.4 m/sec
Group Velocity $C_g = \frac{C}{2}\left(1+\frac{2kh}{\sinh(2kh)}\right)$	= 4.7 m/sec

This latter variable gives the expected in-line translational speed of a coherent packet (i.e., group) in this 8m water depth. Using a mean incident direction of 120 degrees from Figure 7.26 yields the following relative gage spacings for this example:

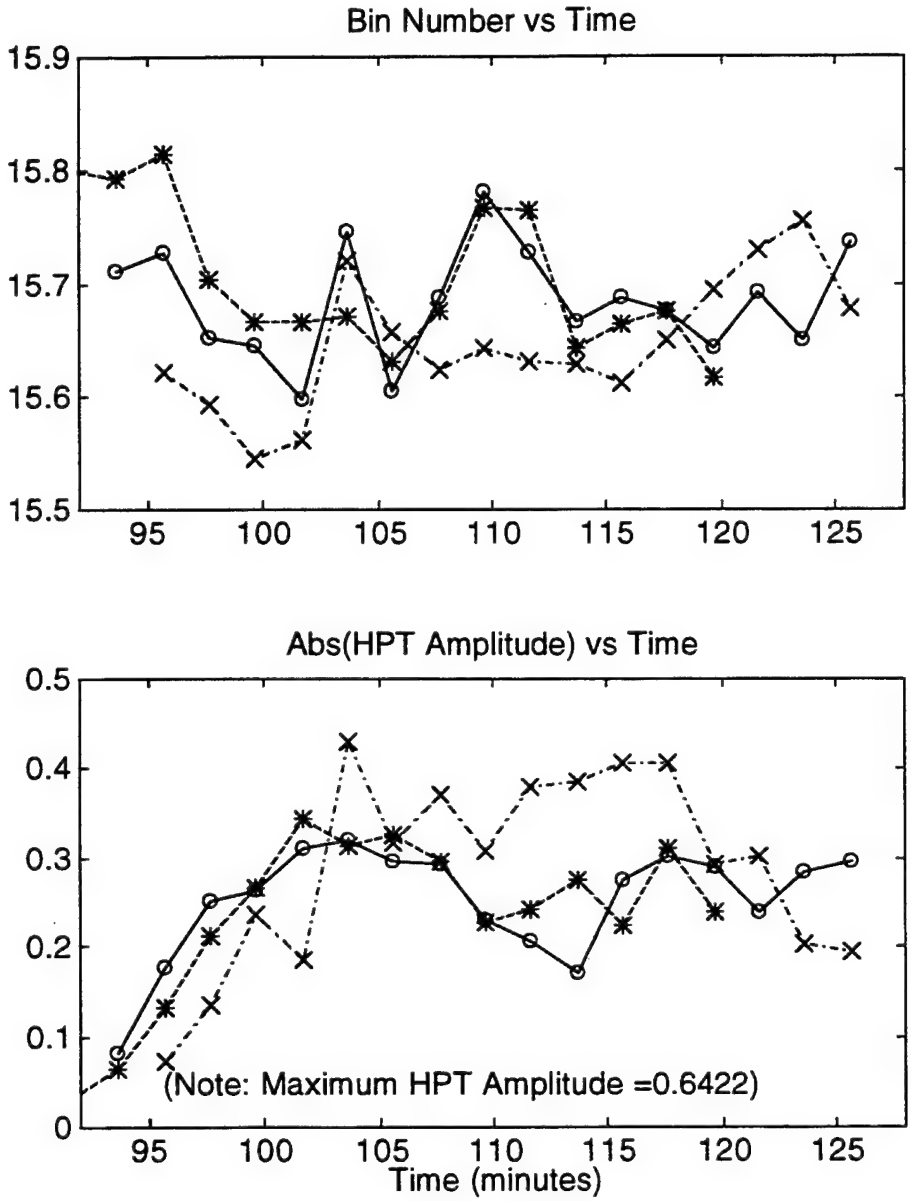

Figure 7.28 Representative Packets During Peak of Hurricane Bob;

—— = Gage 251; - - - = Gage 111, and –·–· = Gage 191

Gage	Relative Spacing (m)	
	In-Line	Orthogonal
111	47.2	1.5
191	185	61

where the most-forward Gage (251) into the seas was used as the reference.

Combining the packet velocity and the in-line spacings yields expected packet time offsets of 5 seconds between Gages 251 and 111 and 20 seconds between Gages 251 and 191. While these time delays are too short to be visible in Figure 7.28, their relative sizes are reflected in the fact that the packet signals for Gages 251 and 111 compare more favorably than the packet for Gage 191.

The correlation among the gages for frequencies above the spectral peak was relatively lower, as it was with the waves before the arrival of the storm (shown in Figure 7.18b). Figure 7.29 illustrates packets over a narrow bin number range for the two closest gages. There is one possibly significant difference, however, between these higher frequency packets before and at the peak of the storm. The duration of the packets in Figure 7.29 are much shorter than their complements in Figure 7.18b. While this could be tangible evidence of the previously stated hypothesis that packets self-organize inversely proportional to frequency, in this case it is more

likely due to wave spreading as the wind direction rotated during the passage of the storm center.

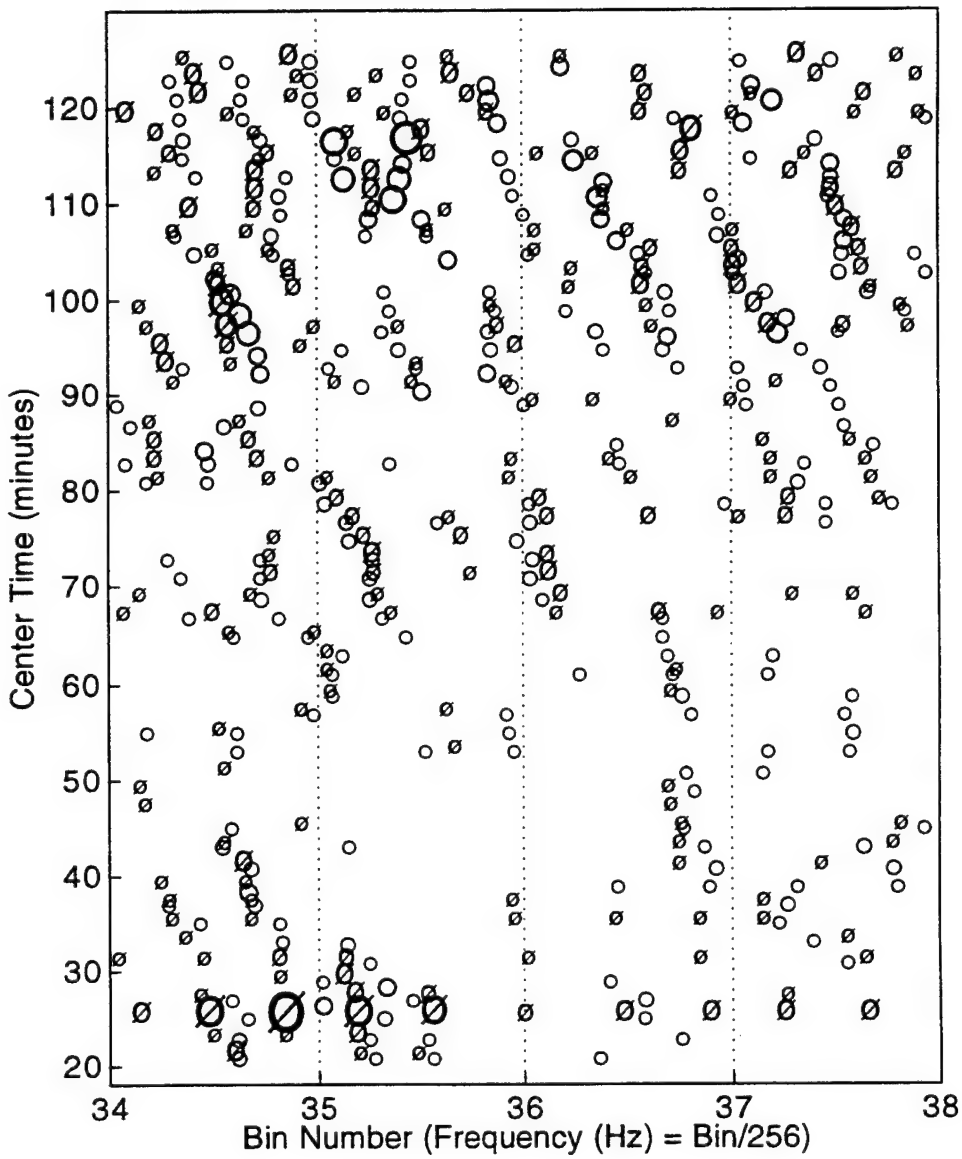

Figure 7.29 Wave Packet Evolution at Higher Frequencies for Gages 251 and 111 near Peak of Hurricane Bob.

Wave packet evolution after the peak of Hurricane Bob is the next topic for investigation. Figure 7.30a, b, and c show evolutions over different frequency ranges for the wave data from the August 19, 0100-0335 block.

An interesting observation from Figures like Figure 7.30a is to track the energy flux, for example at low frequencies. It is clear that wave packets disappear, with no recognizable successor. This issue is most likely related to the large group velocity at low frequencies combined with the finite transit speed of the storm. Figure 7.31 details one of the packets in Figure 7.30a whose amplitude (energy) steadily decreases then apparently disappears. The mean frequency is slightly lower than the mean frequency used previously in Figure 7.28 but the group speed is essentially the same. The two gages are 120m apart (1.4 wavelengths), essentially in-line with the wave direction at this point in the storm, with a maximum correlation coefficient equal to 0.58. Qualitatively, both gages show very similar behavior. The decay in amplitude is very consistent between the two gages, but with a 5 minute relative time lag that is much larger than the lag based on the group speed. These and other differences evident in these figures may be attributed to any number of factors including spreading/short crestedness, reduction due to a new opposing wind direction, and/or numerical error. Further study is warranted before any of these "energy" questions can be answered.

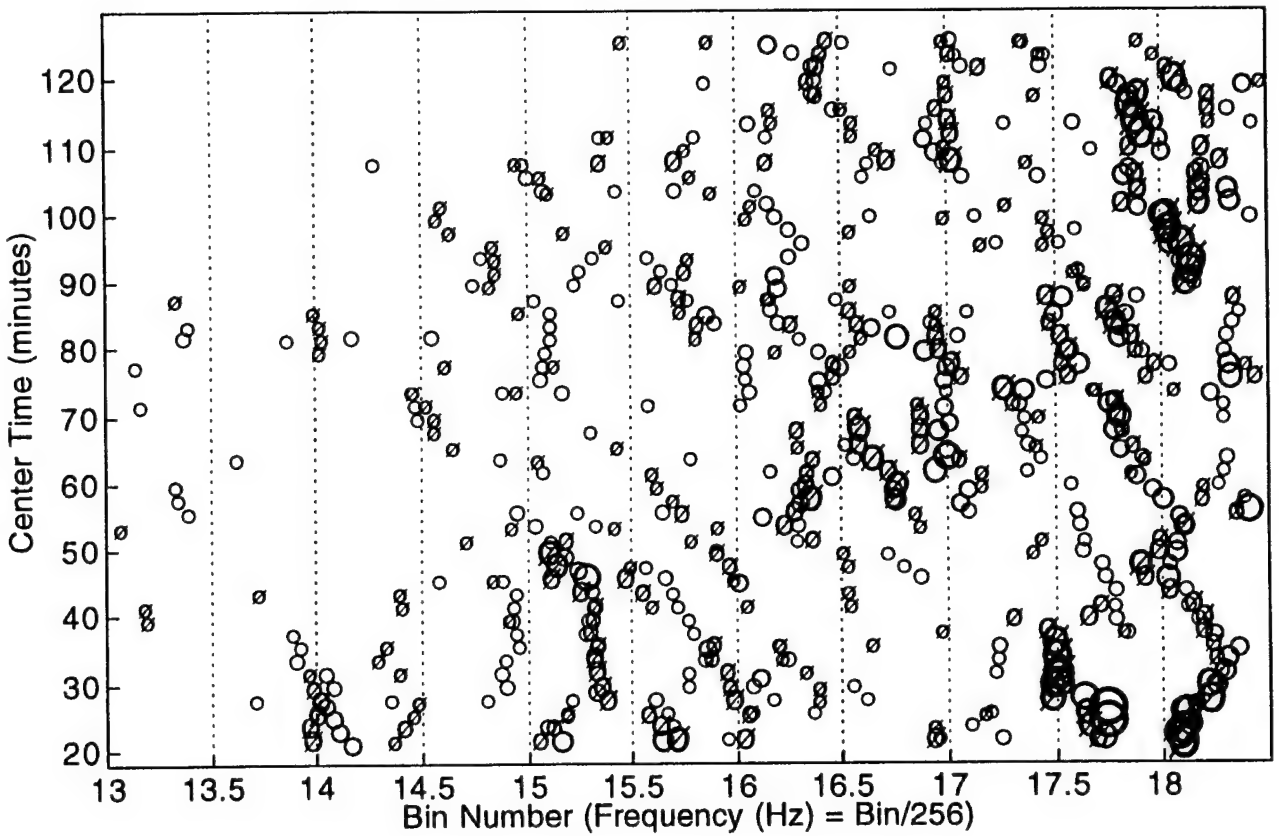

Figure 7.30a Low Frequency Wave Packets for Gages 251 (o) and 211 (ø) after Peak of Hurricane Bob.

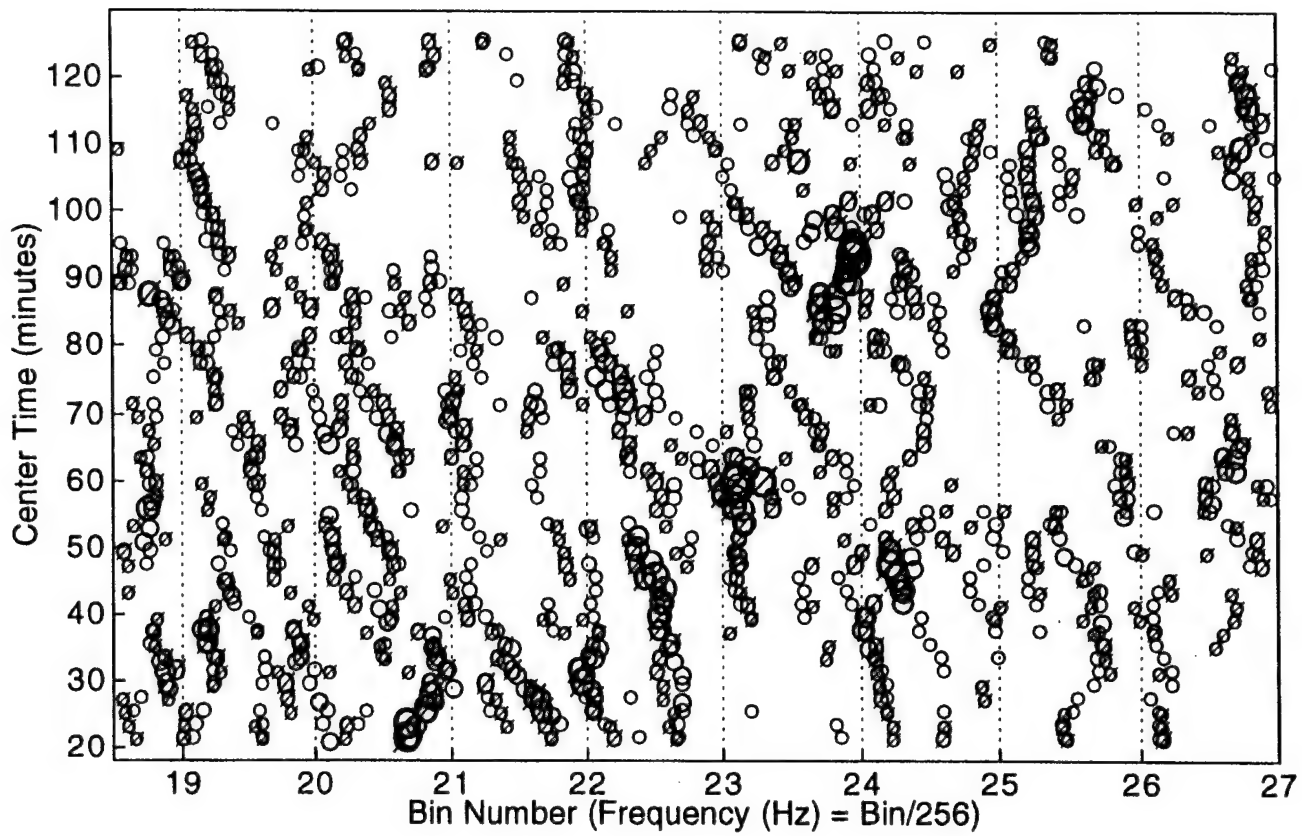

Figure 7.30b Peak Frequency Wave Packets for Gages 251 (o) and 211 (ø) after Peak of Hurricane Bob.

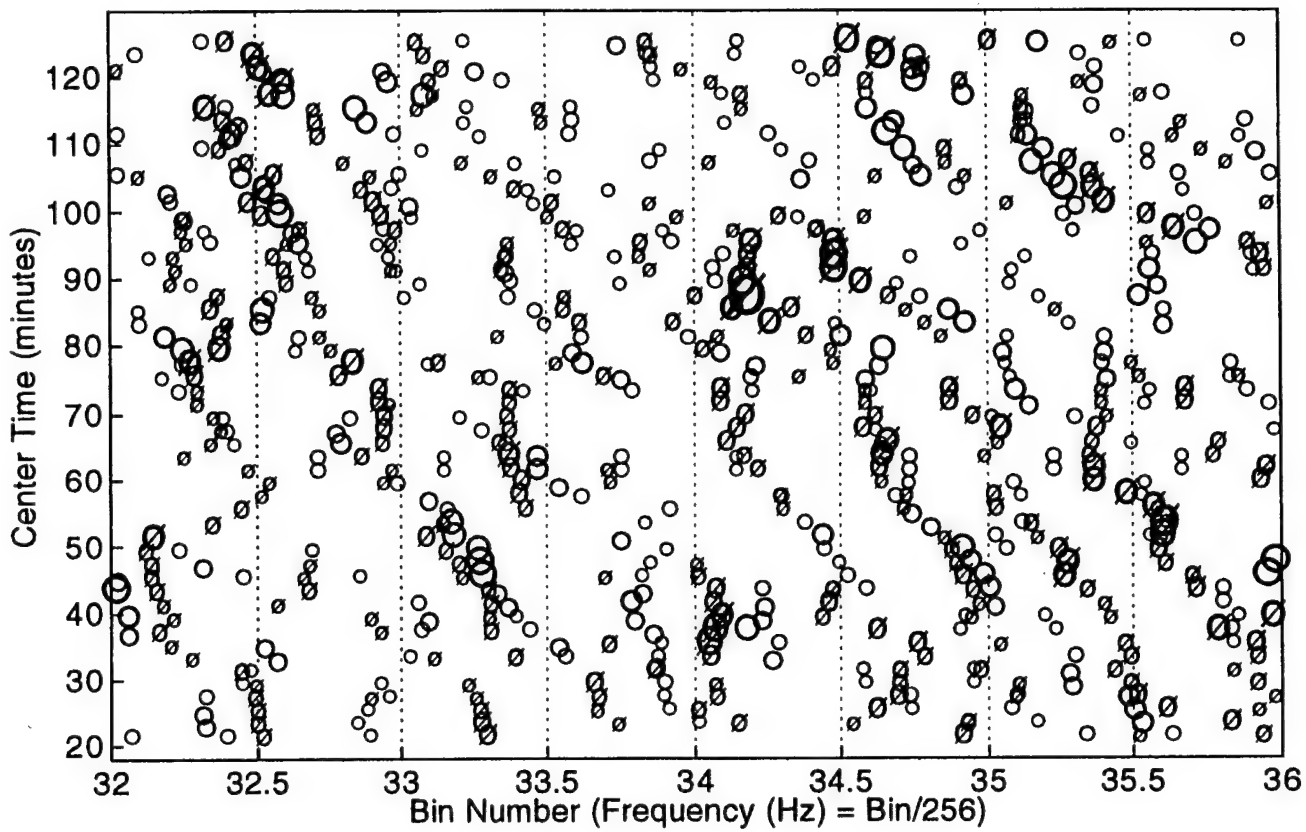

Figure 7.30c Above Peak Frequency Wave Packets for Gages 251 (o) and 211 (ø) after Peak of Hurricane Bob.

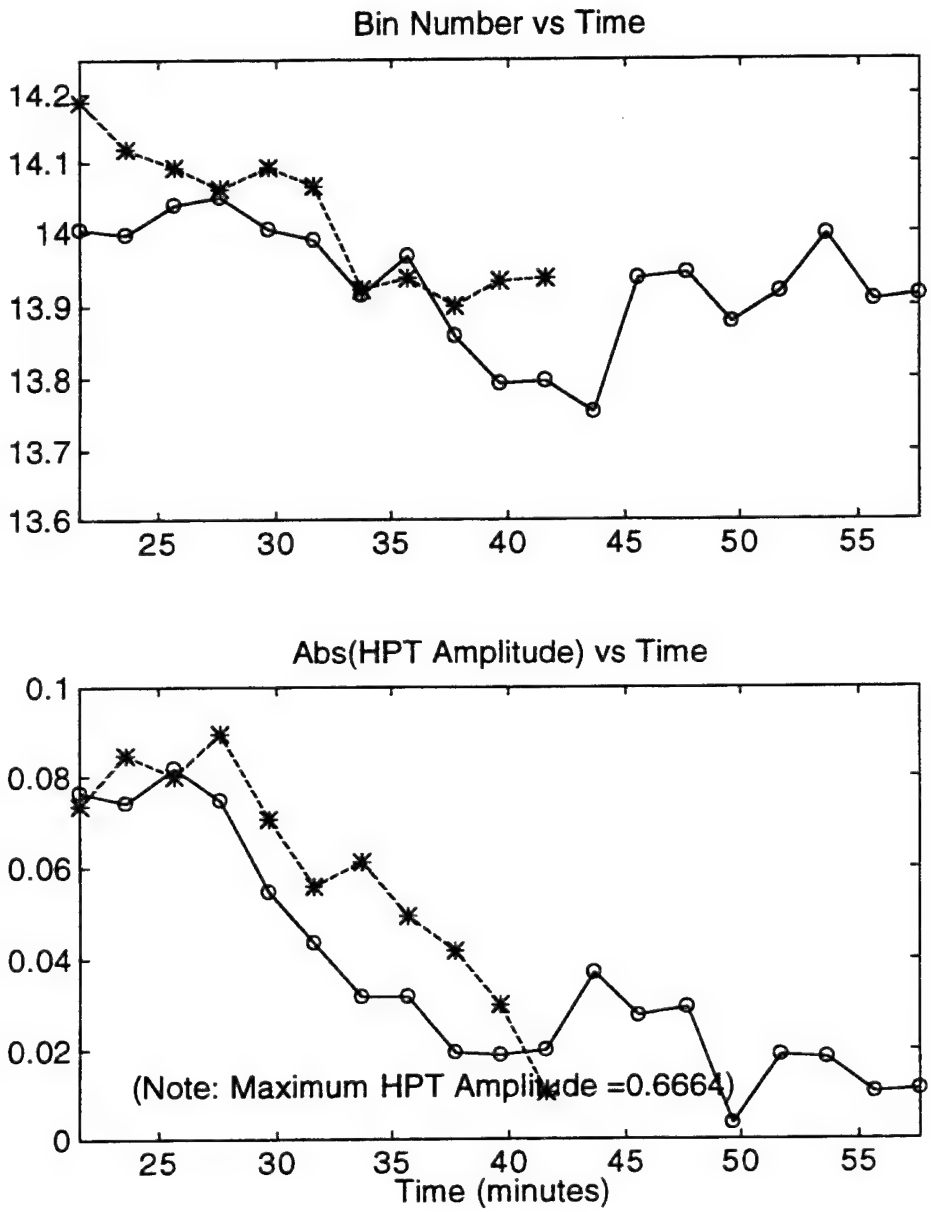

Figure 7.31 Illustration of Disappearing Wave Packet after Peak of Hurricane Bob, Gages 251 and 211.

Figure 7.30b illustrates packets near the spectral peak while the storm was decaying. The downshifting is minimal, and the run lengths vary from short to long. Figure 7.30c illustrates packets above the main spectral peak but centered over the growing secondary peak. The packets in this growing wavefield are seen to consistently shift lower in frequency; run lengths are generally short.

The emphasis up to this point in this subsection has been on the temporal characteristics (evolution) of the wave packets. A short discussion is presented next focusing on the spatial homogeneity of the wave packets.

The HPT information used to estimate wavelengths and directionality contains spatial information from all of the gages over a common time interval. This type of information was illustrated in Figures 7.17 and 7.32 during the build-up of Hurricane Bob. The main indicator of homogeneity in a wave field is invariance, or at least a small deviation that is consistent with the gage positions, of the estimated frequency clusters because that signifies coherent packets. For these clusters where all gages are present and the frequency is relatively constant, then the amplitude and phase can be legitimately inspected to define the characteristics of that particular packet. Conversely, local regions with frequencies that are randomly spaced signify a "confused" sea where the homogeneity needs additional information to resolve. In the figures accompanying this text, the gage estimates are labeled as follows:

North-South Array		East-West Array	
Gage	Label	Gage	Label
191	a	211	i
181	b	221/231	j
171	c	[131]	f
111	d	241	k
121	e	251	l
131	f		
151	g		
161	h		

Table 7.4 Gage Nomenclature for Homogeneity Figures

The reader is referred to Table 7.1 for the relative spacings. The reason for providing Table 7.4 is to allow for spatial interpretation of the upcoming figures.

The stationary wave field of September 13 1990 described in Section 7.3 is investigated first. Figures 7.32a through 7.32c show representative frequency ranges below, across, and above the peak spectral frequency (note the different ordinate scales). Figure 7.32 is essentially an expansion of Figure 7.4b with more gages and a wider frequency range. Observations from these figures are somewhat contrary to initial expectations. The high frequencies, where the wavelengths are shortest

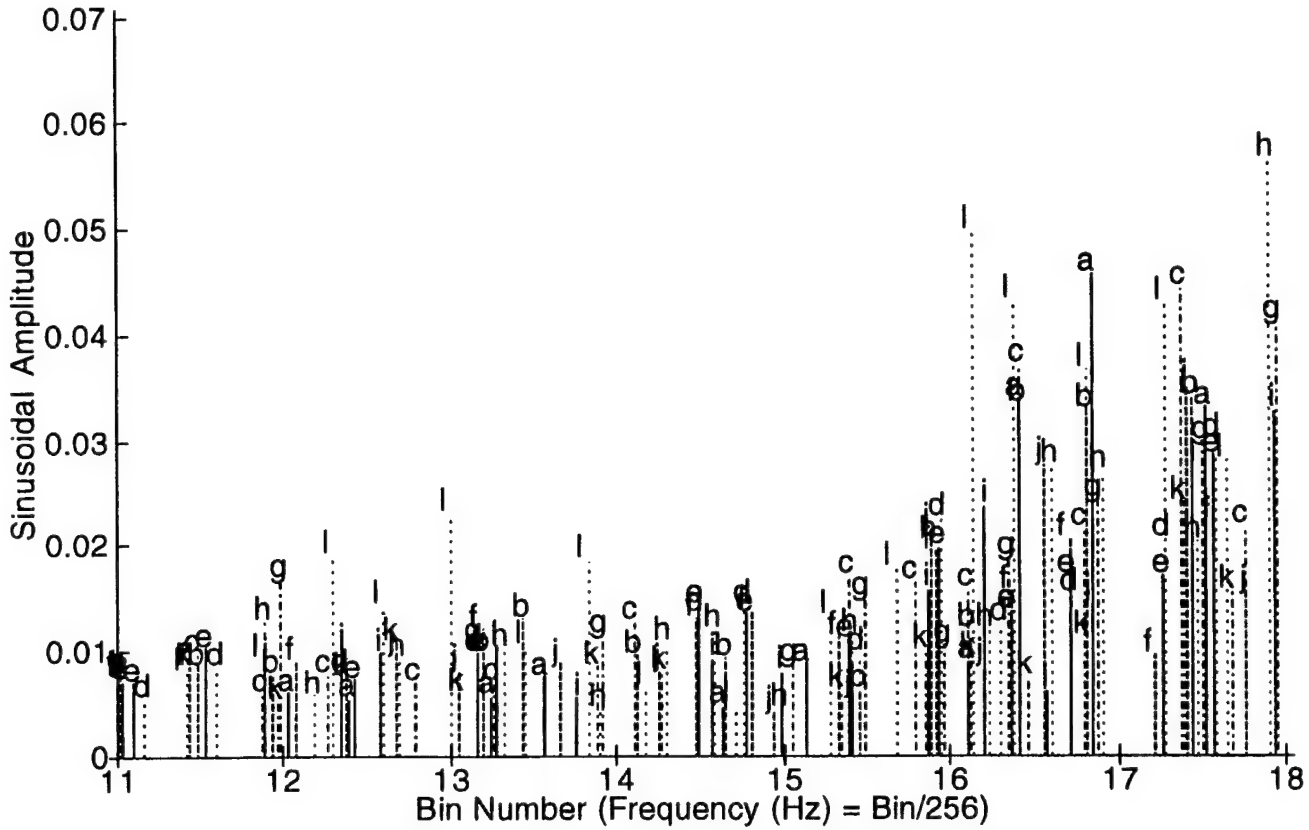

Figure 7.32a Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak in a Stationary Wave field

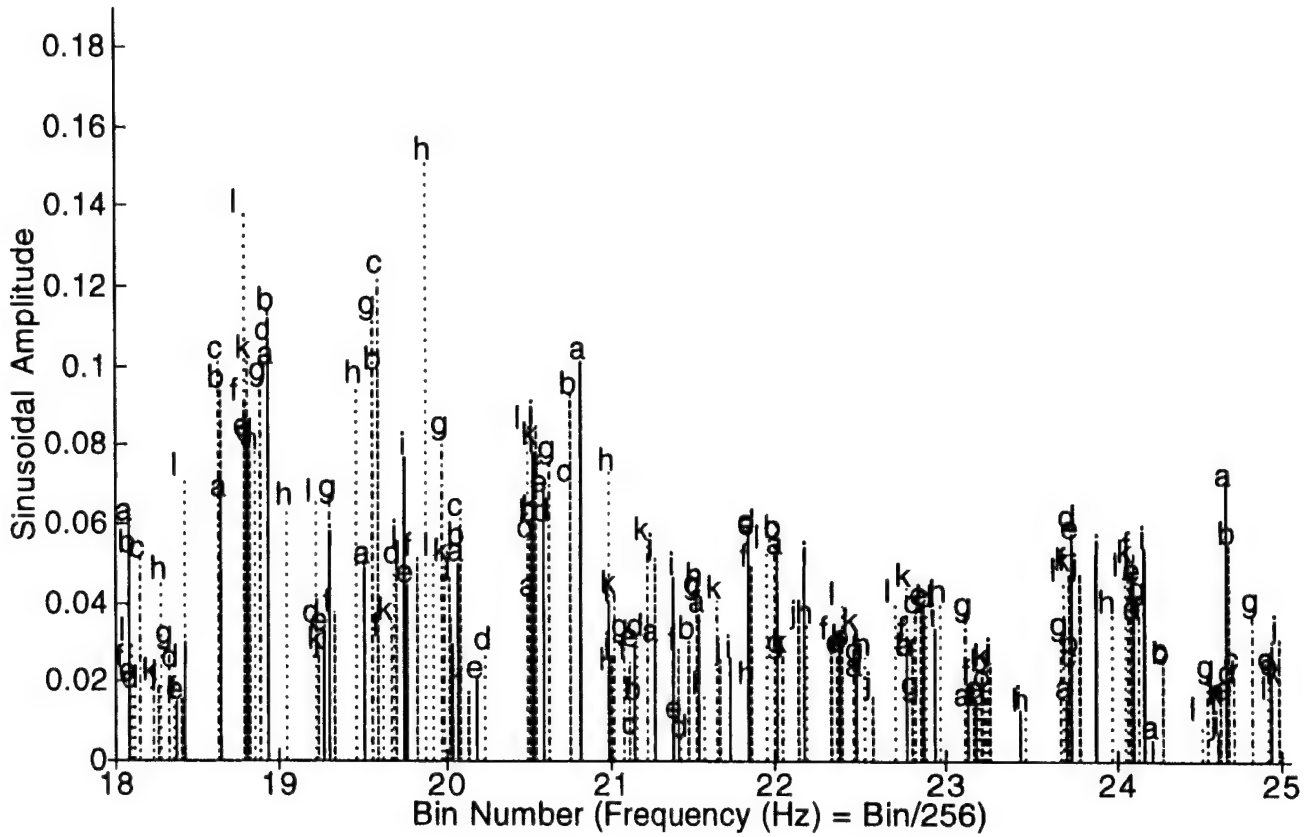

Figure 7.32b Illustration of Spatial Homogeneity for Frequencies Across the Spectral Peak in a Stationary Wave field

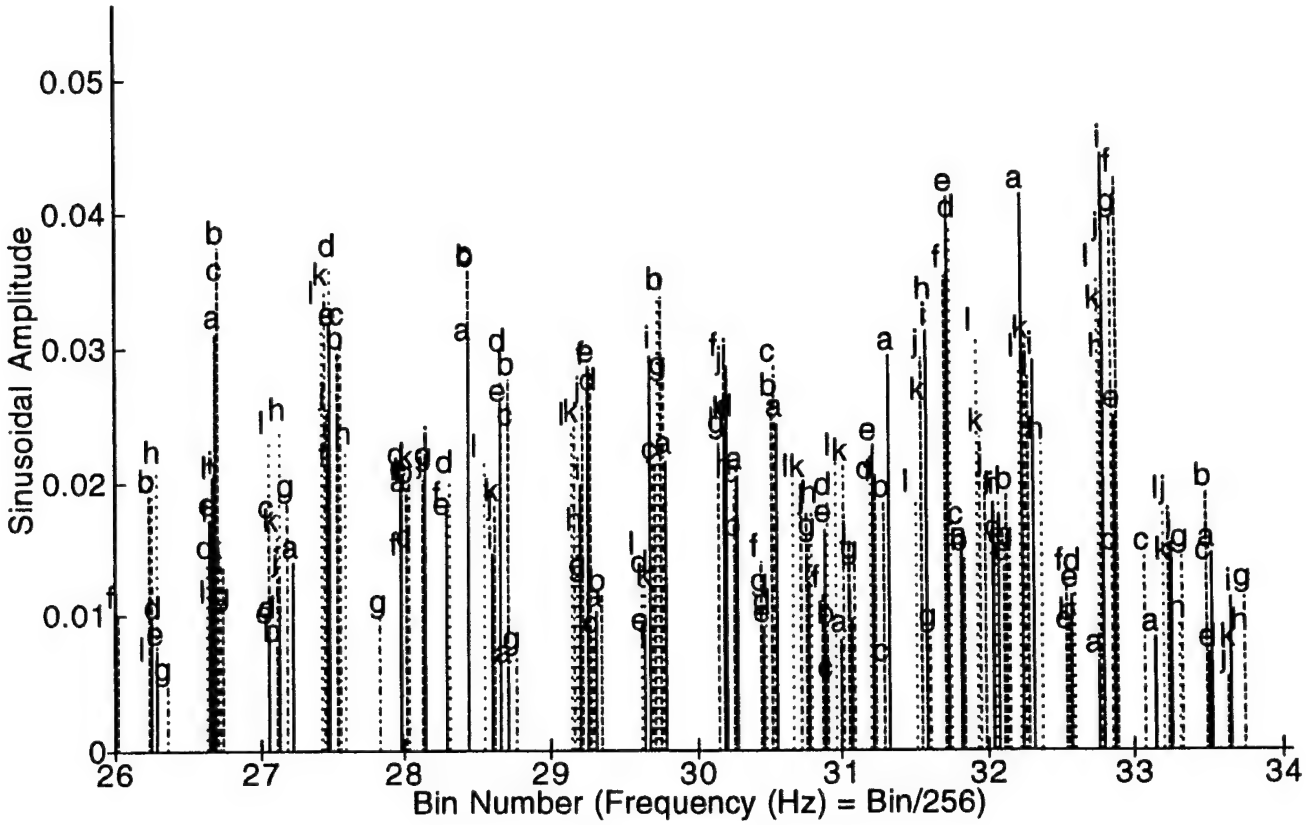

Figure 7.32c Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak in a Stationary Wave field

and therefore the relative array spacing is the highest, show definite clustering which is not as evident over the other frequency ranges. While there are some well-defined packets evident near the peak, on the whole this "stationary" wave field is not particularly homogeneous, at least not to the degree indicated by the stationarity of the spectra in Figure 7.10.

The remaining figures examine the nonstationary wave field corresponding to Hurricane Bob. It was found that the clustering of these packets were remarkably consistent when categorized into relative frequency ranges.

Figures 7.33a through c illustrate homogeneity during the build-up, near the maximum, and after the maximum of the storm but only for frequencies below the respective spectral peaks. In all three cases the clustering is relatively good for this relatively-young wave field, whereas the clustering was poor for the "old" stationary wave field example.

Figures 7.34a through c illustrate homogeneity during the build-up, near the maximum, and after the maximum of the storm for frequencies that straddle the respective spectral peaks. In general, the clustering at the spectral peak is poor, from which it is inferred that there is a large degree of variability in those particular packets. Figure 7.35 details two clusters from Figure 7.34a that further illustrate the potential information available from HPT. The arrows indicate the consistent progression of the

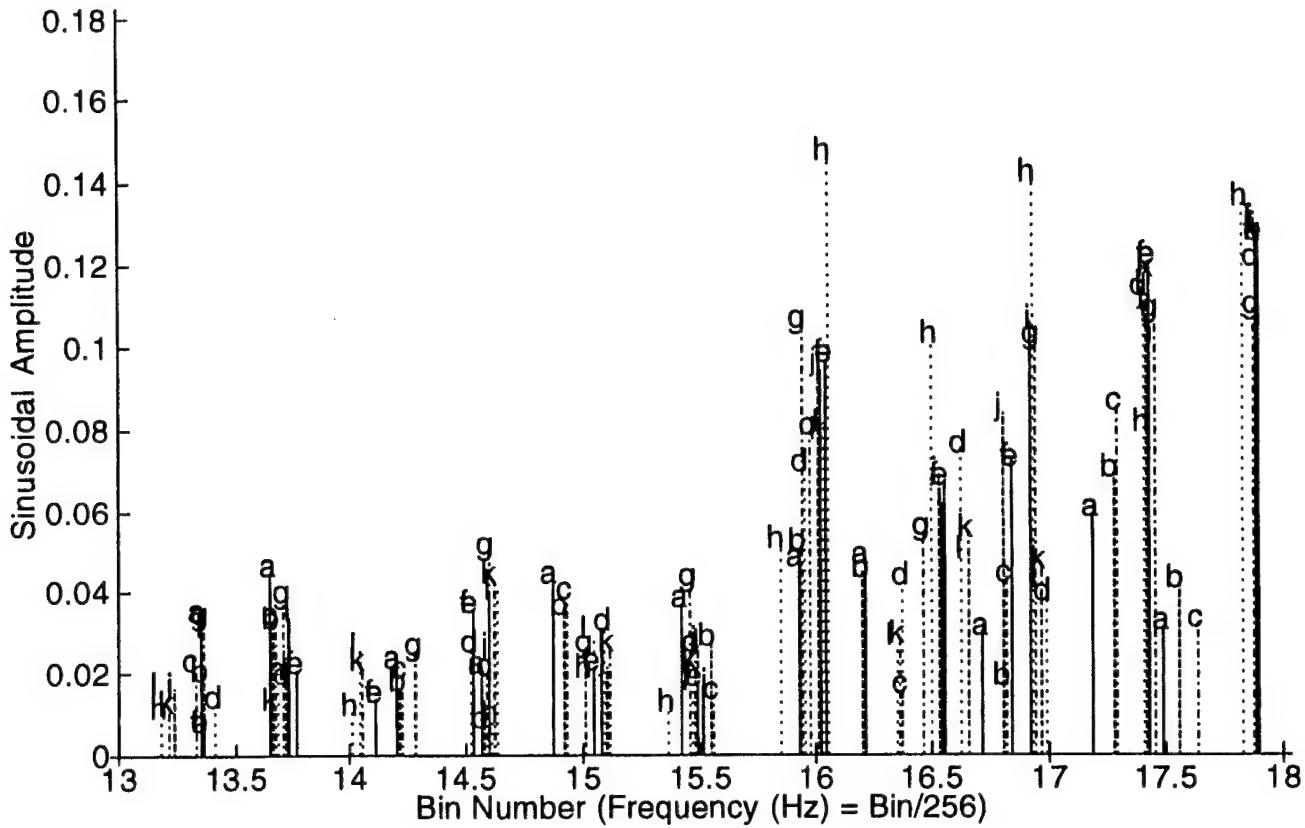

Figure 7.33a Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak During the Initial Stages of Hurricane Bob.

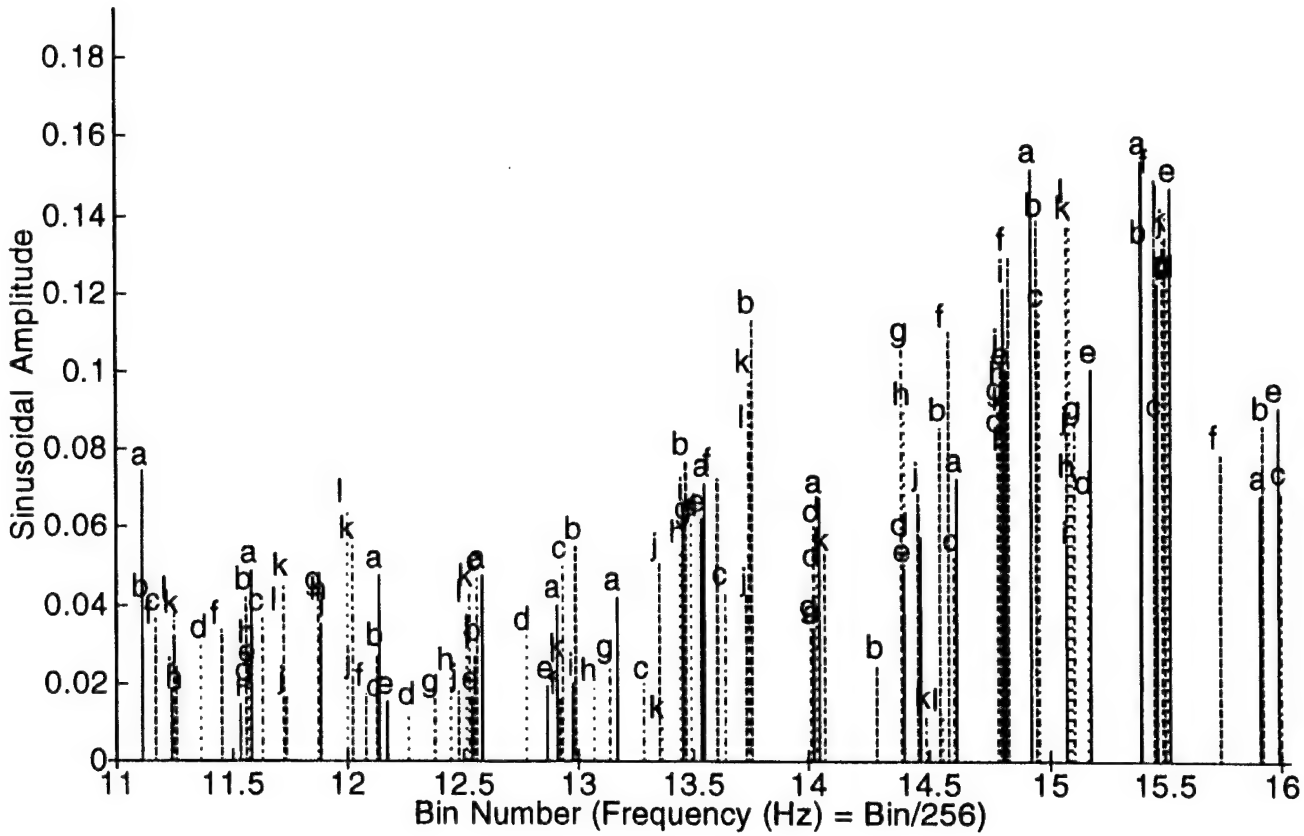

Figure 7.33b Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak During the Maximum Stages of Hurricane Bob.

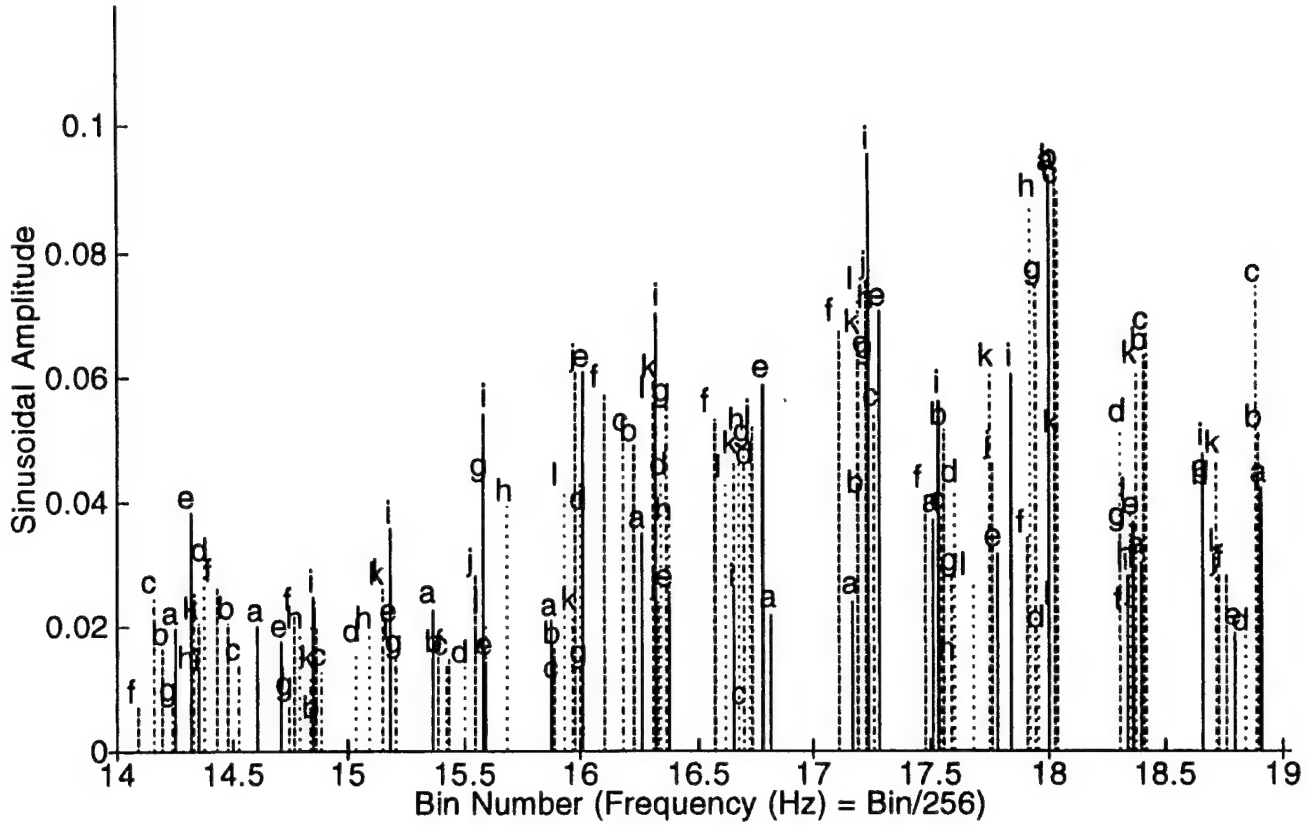

Figure 7.33c Illustration of Spatial Homogeneity for Frequencies below the Spectral Peak After the Maximum Stages of Hurricane Bob.

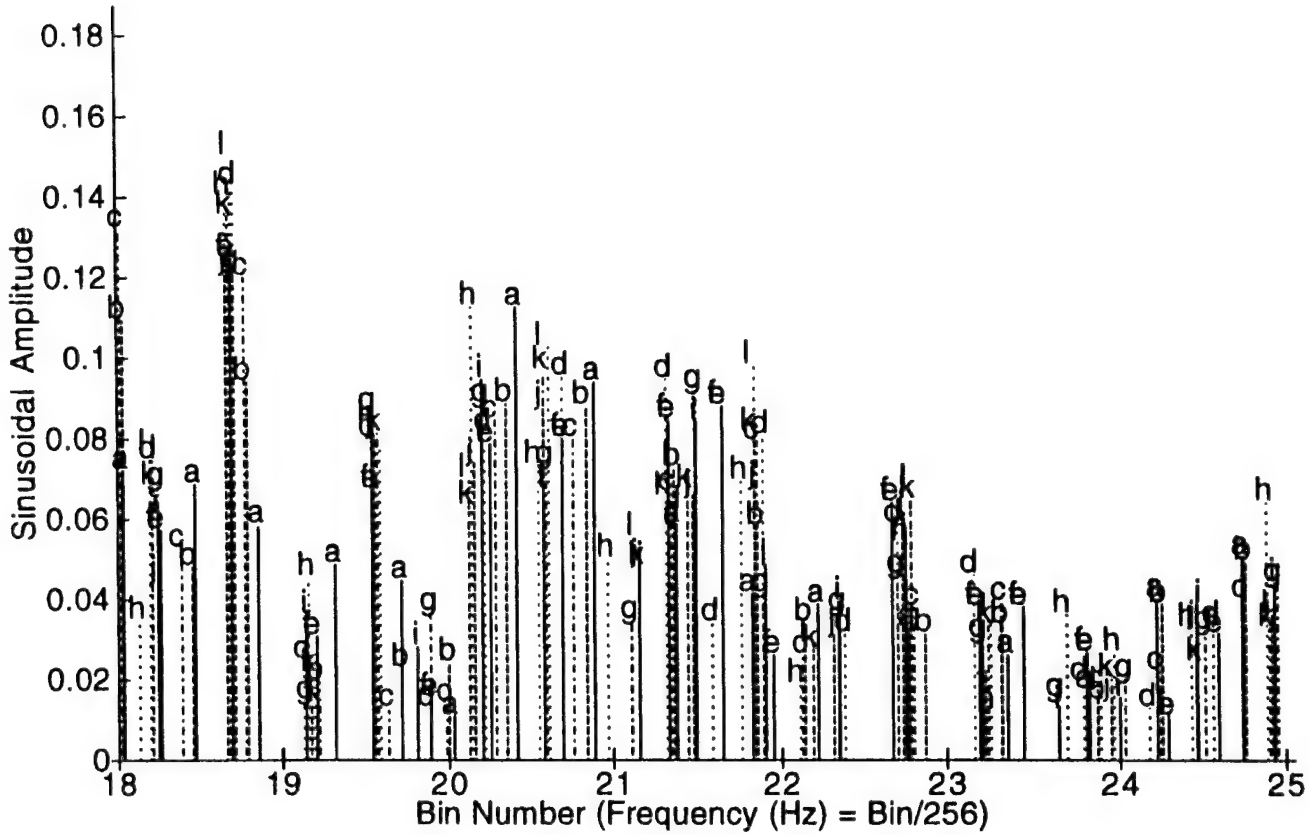

Figure 7.34a Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak During the Initial Stages of Hurricane Bob.

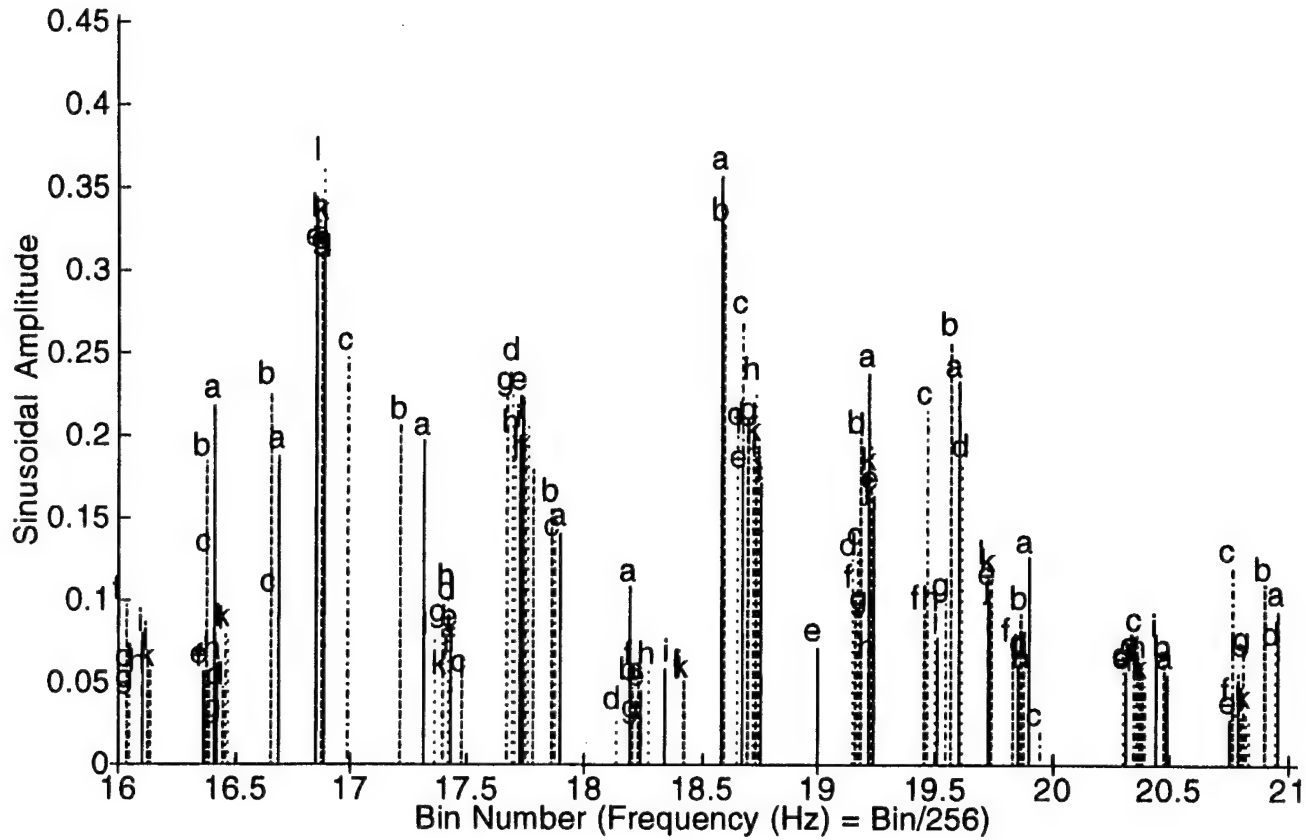

Figure 7.34b Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak During the Maximum Stages of Hurricane Bob.

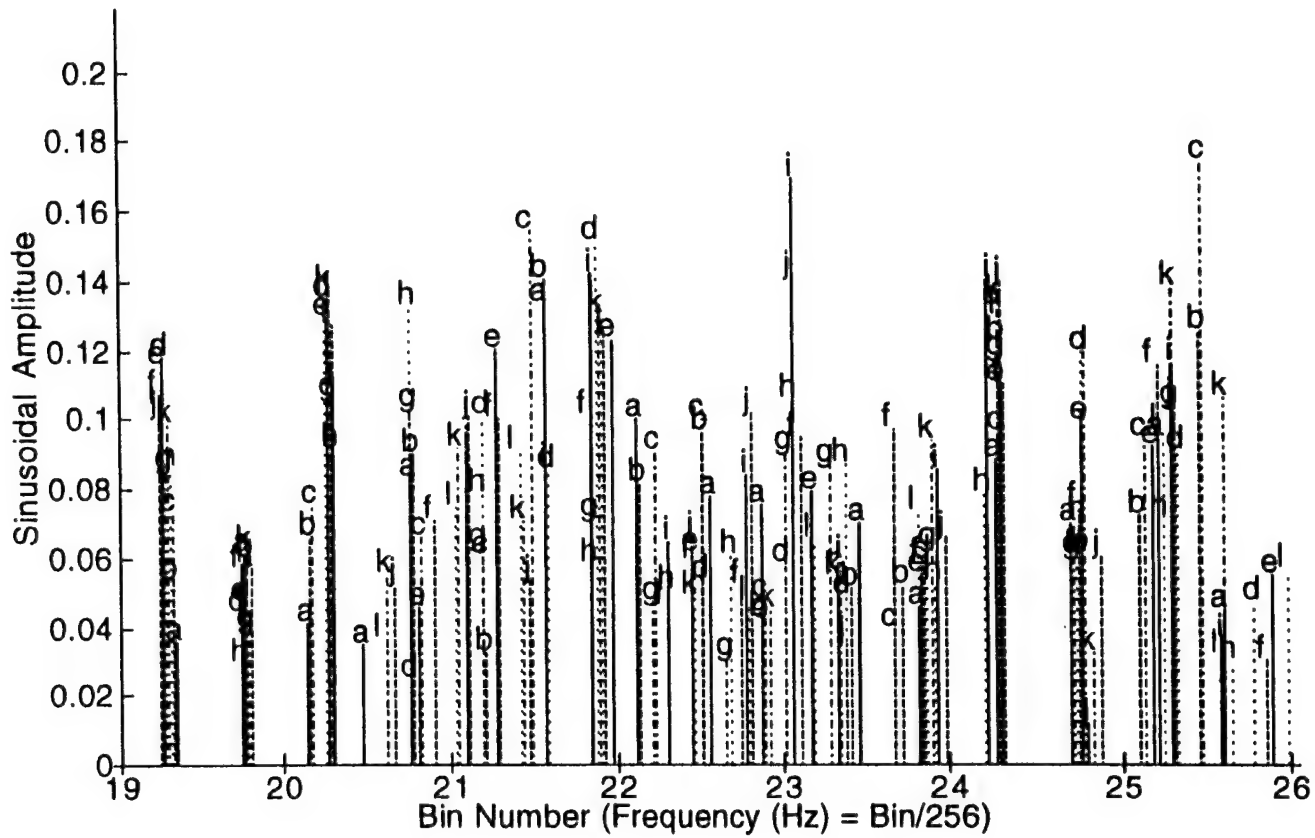

Figure 7.34c Illustration of Spatial Homogeneity for Frequencies that Straddle the Spectral Peak After the Maximum Stages of Hurricane Bob.

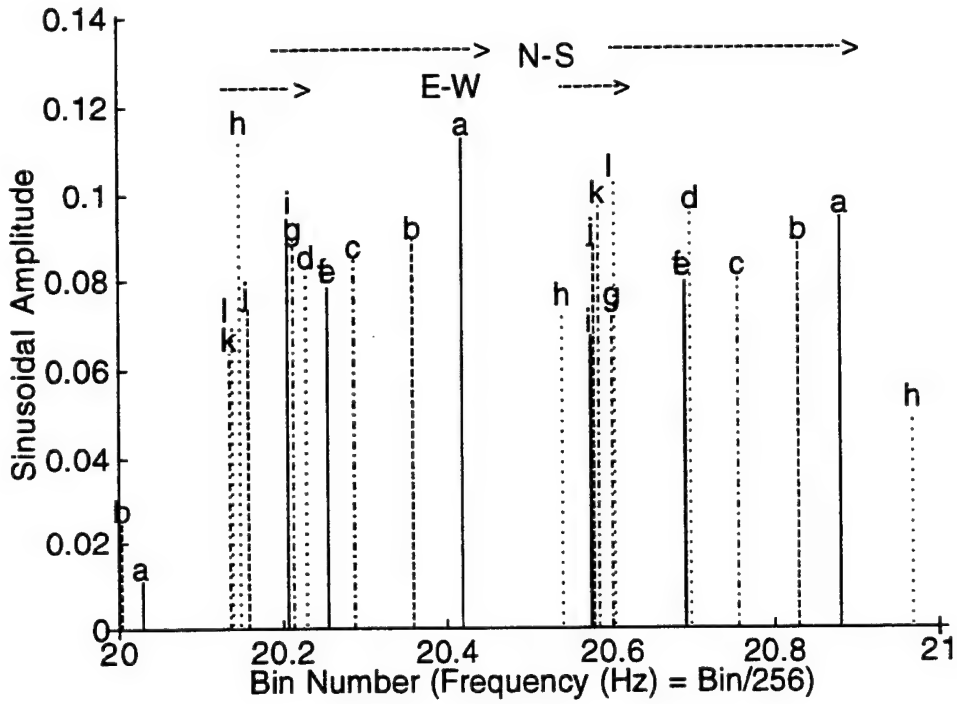

Figure 7.35 Example Clusters Showing Spatial Homogeneity

estimated packet frequency with respect to the North-South (a - h) and East-West (i - l) gage positions. Given the array configuration and the incident direction from the southwest, these observations define an *up shifting* frequency trend for both packets. While this level of accuracy is not always possible, when it is it provides very useful information on packet homogeneity.

Finally, Figures 7.36a through c illustrate homogeneity during the build-up, near the maximum, and after the maximum of the storm for frequencies above the respective spectral peaks. (The frequency range

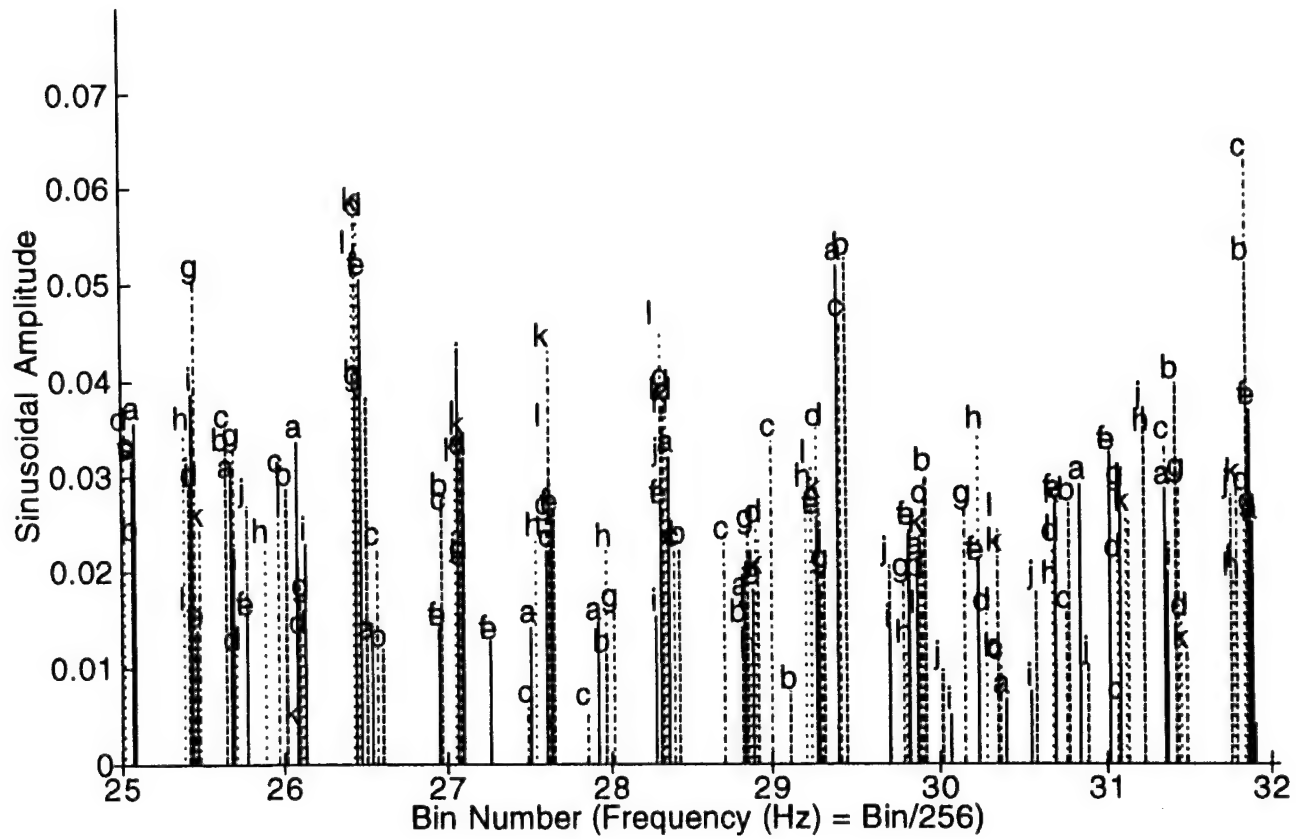

Figure 7.36a Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak During the Initial Stages of Hurricane Bob.

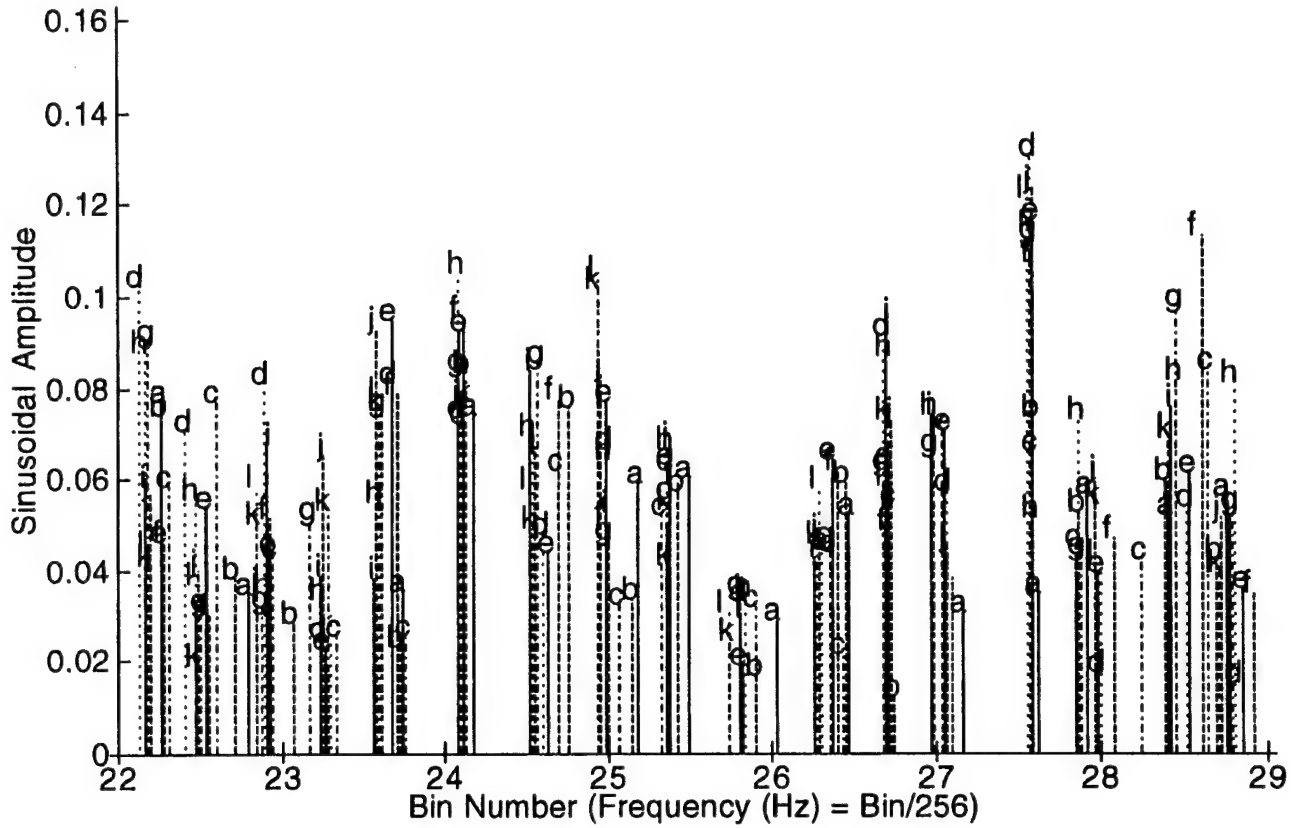

Figure 7.36b Illustration of Spatial Homogeneity for Frequencies above the Spectral Peak During the Maximum Stages of Hurricane Bob.

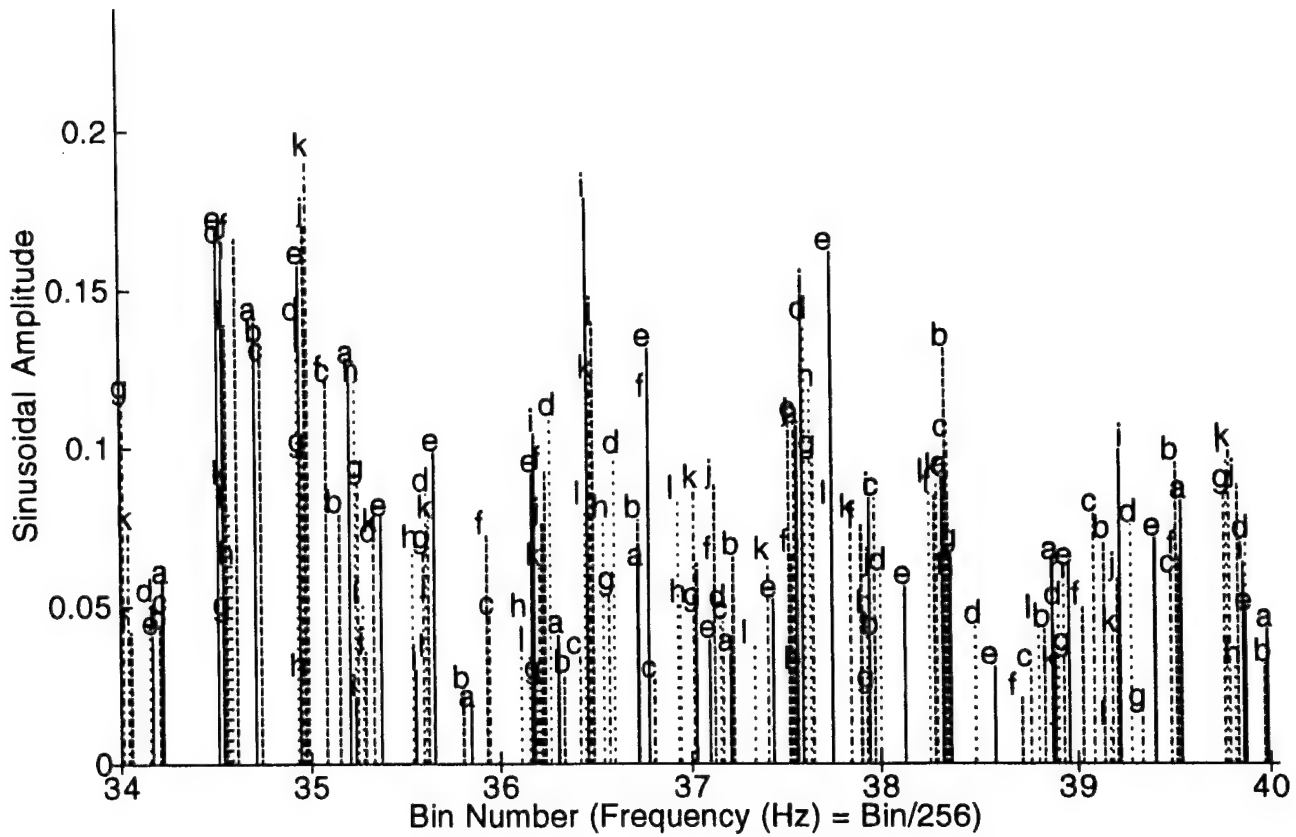

Figure 7.36c Illustration of Spatial Homogeneity for Frequencies above the Second Spectral Peak After the Maximum Stages of Hurricane Bob.

for the waves after the maximum is above the second spectral peak as shown in Figure 7.10.) Again, the clustering above the spectral peak is relatively good, from which it is inferred that there is a large degree of homogeneity in those particular packets.

7.5 Summary of Wave Field Observations

This Chapter examined two representative ocean wave fields, one stationary and one nonstationary. A key aspect of this study was the availability of wave records over a large spatial domain. Harmonic Phase Tracking analyses provided the first known quantitative estimates of the behavior of waves at an intermediate time scale - specifically, the behavior of wave packets in both space and time.

Some brief conclusions are presented here:

1. Wave fields generally self-organize into discrete, well-behaved wave packets that persist for intervals well beyond the wave period.

2. The mean frequency and amplitude of these packets evolve slowly over time (and therefore space).

3. Wave packets are poorly organized at the spectral peak.

4. Packets can demonstrate both downshifting and up shifting of mean frequency over time.

5. Wave packets were observed to disappear at low frequencies under circumstances such as after the storm peak when, due to the close proximity of the storm center, all of the high celerity waves have dissipated.

6. Pairs of packets can apparently merge into single packets if the mean frequencies come together. This behavior is evident in many of the evolution figures.

7. Packets were very organized with respect to space during the nonstationary storm, compared to the poor organization evident in the stationary waves. This may be related to the "age" of the wave field.

It was not possible to make definite conclusions because of the number of variables and different environmental conditions. A very comprehensive study would be required first to quantify how wave packets behave relative to components with differing frequencies and collinear and non-collinear incident directions.

CHAPTER 8

HARMONIC PHASE TRACKING: A PERSPECTIVE

8.1 Chapter Introduction

The emphasis of the previous Chapters was to develop and then very deliberately demonstrate the capabilities that Harmonic Phase Tracking (HPT) brings to the field of signal processing. There were some necessary topics, such as uncertainty and interpretation, that were deliberately avoided as distractive to those HPT demonstrations. Those topics are addressed in this Chapter.

The first section evaluates and contrasts HPT versus other techniques from a signal processing perspective. The second section discusses engineering applications.

8.2 Signal Processing Issues with Harmonic Phase Tracking

This section has three subsections. The first subsection addresses interpretation of raw and averaged HPT estimates. The second subsection focuses on random error issues, while the third addresses the linear algebra aspects of HPT.

One of the reasons there is such a variety of signal processing techniques is the diversity of analysis objectives and constraints among the applications. No one technique, including HPT, is optimum for all applications, and all the techniques add value in their own way. In some instances this Chapter emphasizes weaknesses in other techniques to better illustrate positive features and the potential of HPT. This is done with a full understanding and appreciation for the overall strengths of those techniques.

8.2.1 Interpretation of HPT Estimates.

This section begins with a review of signals, which forms the framework for the subsequent interpretation and quantification of HPT results. Authors categorize signals in a variety of ways such as deterministic versus stochastic, stationary (to various orders) versus nonstationary, continuous versus discrete spectral content, or with or without *a priori* knowledge of particular signal attributes such as rank or probability

distribution. These categorizations are important because they greatly influence the choice of analysis technique and how the numerical results from that technique are properly interpreted.

The selection of a signal category should always be dictated by the physics of the application, but in practice it is sometimes conveniently defined to justify the most readily-available analysis technique. Unfortunately, there is a high degree of subjectivity inherent in either process. For example, an ocean engineer simulating a dynamic system such as a moored vessel might want to quantify an ocean wave signal over a time scale on the order of one half to one hour while an oceanographer might be interested in describing the *same* waves over the course of an entire storm (e.g., days).

This particular illustration of ocean waves highlights a fundamental problem in signal analysis and the subsequent mathematical handling of that problem. First, it is mathematically expedient in some cases to model ocean waves as "an infinite superposition of waves that are infinitesimally close together in frequency, with infinitesimal amplitudes, arriving over a span of compass directions infinitely close together." This yields a continuous, multi-dimensional spectrum. At the other extreme, the method of choice for engineering purposes is to model one finite segment of ocean waves as a finite sum of orthogonal harmonics - precisely, a one-dimensional, discrete Fourier Series representation. These are two

seemingly incompatible models: continuous versus discrete spectra; and based on either assumed or artificial (falsely periodic) component waves.

Now, consider this incompatibility from the perspective offered by HPT. Define the waves $\eta(t)$ at a fixed position based on the usual infinite superposition model, simplified for one direction for convenience:

$$\begin{aligned}\eta(t) &\equiv \int_0^\infty a(f)\cos(2\pi f t+\theta(f))df \\ &\equiv \sum_{k=1}^{\infty} a_k \cos(2\pi f_k t+\theta_k)\end{aligned} \qquad 8.1a,b$$

where the component frequencies f_k are assumed to correspond to physically-present waves which are not necessarily uniformly distributed and the a_k amplitude includes the (non uniform) df_k. Rearrange the summation in terms of adjacent pairs of components:

$$\begin{aligned}\eta(t) &\equiv \sum_{k=1}^{\infty}\left[a_{2(k-1)}\cos(2\pi f_{2(k-1)}t+\theta_{2(k-1)})+a_{2k}\cos(2\pi f_{2k}t+\theta_{2k})\right] \\ &\equiv \sum_{k=1}^{K} A_k \cos(2\pi f_{envelope}t+\vartheta_k)\cos(2\pi f_{mean}t+\phi_k)\end{aligned} \qquad 8.2a,b$$

where the infinite upper limit has been replaced in the second equation with a finite value by requiring that the signal be realizable with a finite bandwidth. Equation 8.2b shows that a bandlimited wave signal can be alternatively mathematically interpreted as the sum of a *finite* number of modulated sinusoids with no loss of generality.

Note also that any finite number of adjacent components can be grouped and interpreted in this same way, with more and more generality added to the periodic, sinusoidal envelope. Since the frequencies are infinitesimally close, the periods of these envelopes can be very long. Thus, this same model can be interpreted in an engineering sense to model the evolution of wave energy in the various frequency bands at any desired time scale - over one storm, a season, or a century.

The interpretation dilemma arises from the fact that it will never be possible to know the true components, because the measurement time interval cannot be made infinitely long to allow for identification of the "true" infinitesimal components. In that sense, it may be inappropriate to even define "true" components. *Regardless, all subsequent estimated signal models based on finite wave records are approximations.* Fortunately, this open-ended time scale is often bounded for geophysical signals (e.g., ocean storm waves, earthquakes) which have recognizable starts and stops that are used to define the maximum length of a coherent event; this is certainly the case for man-made signals such as radar.

A situation where the interpretation problem is not as applicable is when orthogonal models are used to quantify the signal. These estimates have a number of well-known drawbacks associated with: effects due to the fundamental mapping of the original signal into a periodic approximation, leakage whenever the frequency of a physical component wave does not

match one of the *a priori* mathematically convenient orthogonal periods (equal to integer divisors of the record length for FFTs), and the need for ensemble averaging, typically accomplished by assuming ergodicity and performing time rather than ensemble averages. Since it is impossible with these orthogonal techniques to ever model the physical components (because making the frequency resolution small necessarily pushes the limits of stationarity of a signal), the question of physical interpretation of this representation and hence the dilemma were both often avoided. But consider instead results from an HPT analysis. Instead of averaging energy over finite frequency bands and then forcing the average to a discrete orthogonal frequency set, HPT more accurately estimates "best" frequencies within a given band (and with slightly better resolution). These frequencies are equal to localized instantaneous frequencies if the spectrum is continuous, but correspond to the signal component frequencies when the spectrum is discrete (finite rank signal) and they are spaced wider than the resolution limit. [Aside: it is suspected that the ability of HPT to find "the best/minimum rank" frequency vector could be of interest regarding the issue of parsimony for some signal types.]

Therefore, regardless of the signal spectrum, the HPT frequencies are associated with the underlying physical component(s). But there is still an apparent problem - HPT is capable of reliably identifying the underlying components only when the segment length is greater than half the modulating period. Since the modulating period nor whether an identified modulation is only one part of a longer modulating cycle are not

a priori known, the analyst cannot be sure based on one analysis of whether a HPT component represents a fundamental signal component or an effective instantaneous [beating] component.

As with other techniques, HPT will estimate different components versus different trial segment lengths. But whereas estimates based on orthogonal sets converge in a *mathematically* consistent sense, HPT estimates converge in a *physically* consistent sense. This is a fundamentally different concept that can be used to better understand the signal. The connection between HPT and the alternative interpretation of a continuous spectrum in Equation 8.2b as a finite sum of modulated component signals can now be appreciated. Recall the examples in Chapter 5 that demonstrate how HPT can model many types of signals including frequency- and amplitude-modulated component signals. *The conclusion is that, while HPT uses constant parameter sinusoids to model a signal, the signal components are not limited to that restrictive class.* These trial HPT frequencies, amplitudes, and phases versus segment length should, ideally speaking, follow a "converging trend", and it is the examination of this converging trend that is proposed as unique to HPT. Specifically, recall the many discussions and examples in the main text (particularly Chapter 5) and Appendix A regarding the relationship between the instantaneous frequency and the sum of two (or more) closely-spaced harmonics. As the trial HPT segment length is reduced, distinct adjacent components should merge, with attributes corresponding to the local instantaneous frequency. Furthermore, these attributes

should be quantifiable by examining the continuity of the phase signal across adjacent segments (or conversely, the lack of continuity as illustrated by the analysis of white noise signal in Figure 5.15b). The fact that these HPT estimates have a consistent "physical" interpretation might be used to great advantage because it potentially provides a continuous link for estimates over all frequency resolutions.

Inspection of the phase continuity information from HPT over a series of estimates could be used to set an upper bound on the time interval of stationarity for a given component, and in turn set a lower bound on the resolvability of the "true" components in the vicinity of that frequency. This would also answer the question of whether a spectrum was continuous or discrete. The behavior of this time interval versus frequency might be fundamentally different for a geophysical signal compared to a filtered version of this signal passing through a low damped dynamic system with multiple resonances. This is recommended as a rich topic for future study.

If HPT does allow for consistent identification of components over any time scale, then it opens the door for a more quantitative handling of uncertainty for nonstationary signals. Even when a signal is known to be stochastic and nonstationary, when only one measurement is available and time averaging is not possible then with existing practice the *same* signal is redefined as deterministic for subsequent analyses such as wavelets, the Wigner Ville (or similar) distribution, or Short Time Fourier Transform techniques. In other words, error bounds are not as rigorously

required for these signal descriptors. As demonstrated by the numerical examples in Chapters 5, 6, and 7, the plots of HPT "raw" frequencies and amplitudes versus time quantify how individual signal components evolve. It is suggested that a low-order polynomial could be used to convert these "raw" estimates into "averaged" estimates, and in the process provide a formal definition for the averaged estimates with an associated regression-based measure of variance. Note that these functions would be valid *independent of the stationarity of the components.* This is another suggested topic for future research.

8.2.2 HPT Random Error Issues.

None of the numerical examples presented in the previous three Chapters included uncertainty intervals for the "raw" HPT estimates. Before addressing them it is instructive to categorize HPT with respect to other model categories:

1. HPT is a high resolution technique. All of these techniques are known to be not linear operators. That makes it very difficult to develop an analytical uncertainty expression.

2. As outlined in Chapter 4, the HPT basis matrix is *random* because it is a function of the unknown coefficient vector:

$$\mathbf{x}_{\mathrm{HPT,even}} \equiv \mathbf{H}(\mathbf{c})\mathbf{c} + \mathbf{n} \qquad 8.3$$

where $\mathbf{x}_{HPT,even}$ is the approximation to the even component of the [Fourier transformed] data, **H** is the least squares basis matrix based on the iterated frequency vector found from the coefficient vector **c** of in-phase component amplitudes, and **n** is the noise vector. (The expression for $\mathbf{x}_{HPT,odd}$ is equivalent.) All of this has implications (not pursued) regarding optimality of the solution.

The question of bias is difficult to quantify for HPT estimates for two reasons. First, no analytical methodology was identified because of the nonlinearity and the iterative nature of HPT; other high resolution techniques have similar problems and typically require Monte Carlo simulations to numerically define the errors. Second, the whole question of bias is intertwined with the concept of "true" component values discussed in the previous subsection. If bias is somehow defined, then it seems that the bias of the HPT raw frequency vector should be treated separately since it is the most important factor influencing the subsequent amplitude and phase parameter estimates from the total least squares analysis. This treatment should also include the averaged estimates found, for example, from a low-order smoothed estimate based on the raw evolutionary plots.

The reader is reminded that the inherent capacity of HPT to adjust the frequency vector means that the HPT bias, while not analytically expressed here, is certainly less than the bias from any orthogonal model

for all signal types. For example, bias in FFT-based spectral ordinates is proportional to both the local second derivative of the ordinate function (with respect to frequency) and the ratio of the frequency resolution divided by the bandwidth of the local spectral peak (Bendat and Piersol, 1993). [As a simple example, the bias in HPT raw amplitude estimates for the ideal case of noise-free, constant parameter, discrete harmonics spaced wider than the frequency resolution limit is presumed to be almost negligible given the behavior of the $\mathrm{sinc}(\delta f)/\delta f$ function for the small frequency bias δf expected from HPT.] The same reference makes the argument that there is a time bias in nonstationary spectral estimates that is proportional to the local second derivative of the ordinate function with respect to time. Lastly, the reader is also reminded of the effect that the spectral window has in modifying the shape (i.e., biasing) of the spectral ordinates.

The variance of the raw HPT estimates is the next uncertainty topic. For general geophysical applications the signal contains a large number of sinusoids. In these cases it is not possible to derive the Cramer-Rao bound because the bias is not known, although there are complicated variance expressions available for "multiple sinusoids" based on numerical evaluations (Kay, 1988). If it is assumed that the HPT frequency estimates are relatively unbiased, and the frequencies are spaced "much farther than $^1/_T$ apart", then further simplifications are possible and the Cramer-Rao bound may be more useful.

A more comprehensive study of bias and variance for both raw and averaged HPT estimates versus signal characteristics (like signal to noise ratio, number of sinusoids, relative frequency spacing, etc.) and segment length was not pursued but is needed; the most likely approach is Monte Carlo numerical studies.

8.2.3 Linear Algebra Issues with HPT

This final short subsection discusses basis, rank, and spaces for HPT modeling.

As a reference, consider the linear algebra aspects associated with the use of the FFT for signal analysis (refer to Section 3.2). It is well known that the span of orthogonal sinusoids with integer harmonics of the record length used as the invariant FFT basis matrix is sufficient to fit most real-world signals (except for discontinuities such as found at the ends due to the assumed periodicity). The rank of this matrix is generally N, where N is the number of points in the data vector. Furthermore, the condition number of the (diagonal) matrix is always optimum at 1.0.

In contrast, the basis matrix for HPT is not predetermined but instead adapts during the fitting. Similar to the FFT, HPT does span almost all possible signals since the HPT frequencies can range continuously

between $0.4/(L*\Delta t)$ (lower limit used in this study) and the Nyquist frequency equal to $^1/_2/_{\Delta t}$, where L is the number of points in the HPT segment. Admittedly, it does not model some signals, such as the even residual signal associated with a sinusoid with a linearly-varying amplitude as discussed in Section 5.2.3. But the big advantage is that, in most cases, HPT finds the minimal rank model necessary to fit the signal, unlike the FFT which in general applications requires a rank N model to fit a length N data vector. The in- and out-of-phase HPT basis matrices before the addition of the right hand side data vector during the total least squares solution process (refer to Equation 4.30) are generally fully populated and have full rank (i.e., the sinusoids corresponding to the HPT frequencies are dependent). The total least squares solution is straightforward when the right hand side vector includes noise; since that is almost always the case with measured geophysical data there is never a problem. The reader is referred to Van Huffed and Vandewalle (1991) for a comprehensive treatment of total least squares issues. Also, the condition number of the HPT total least squares basis matrices are typically between 3 and 10, even for matrices on the order of 100 by 100.

The data vector used for the previous paragraph was purposely labeled *signal*. There is a fundamental reason for this. An orthogonal model like the FFT is numerically very efficient so it can be used to fit coefficients to all of the frequencies. On the other hand, HPT (as implemented here) is not as efficient; the matrices are fully populated, the technique is

iterative, and each iteration requires multiple singular value decompositions. Thus, HPT computational time is proportional to the size of the matrices squared, the number of iterations, and the number of time shifts per iteration. While the latter factor was arbitrarily fixed for these studies, the user can minimize the first two factors by adjusting the amount of error tolerable in the final fit. This acts to minimize the number of components in the matrices and reduce the number of iterations during the asymptotically-convergent iterations. For example, each new HPT component sinusoid increases the size of the basis matrices, yet in many cases these sinusoids correspond to low energy noise that does not significantly contribute to the overall fit. As a consequence, it can be prudent for the analyst to define HPT solution parameters such that these components are excluded from the HPT fit. While this compromise between the computational time and the definition of the signal (versus remaining noise) is an admittedly subjective step, in an engineering sense it can be readily defended. Furthermore, if the HPT algorithms were adapted to only include components with significant amplitudes, this would act to minimize the computational time in engineering applications where that was a crucial factor (discussed in the next section).

This compromise in turn defines the rank of the signal. This study did not seriously address the issue of the signal versus noise subspaces and whether they were orthogonal or not. The limited number of inner products examined were not conclusive; for example, signal-to-noise

correlation coefficients of up to 0.10 were found for some of the ocean wave records. Pending future studies, it is hypothesized that these HPT subspaces are not strictly orthogonal, by simple recognition that both subspaces (even white noise) have HPT decompositions comprised of harmonics at arbitrary frequencies that are by definition correlated.

As discussed in the previous subsection, HPT is not a linear operator. The previously demonstrated fact that HPT can identify either one or two sinusoids ($\mathbf{v}_1, \mathbf{v}_2$) depending on their frequency spacing relative to the resolution establishes that the HPT operator ($\mathcal{H}$) is not universally additive between iterations. Thus,

$$\lim_{\delta f \to 0} \left[\mathcal{H}\left(\mathbf{v}_1(f_o) + \mathbf{v}_2(f_o + \delta f)\right)\right] \neq \mathcal{H}(\mathbf{v}_1(f_o)) + \mathcal{H}(\mathbf{v}_2(f_o + \delta f)) \qquad 8.4$$

However, HPT is a linear operator regarding the forward and backward time shift estimates at each iteration because the frequency vector is held constant.

A final linear algebra topic recommended as worthy of study is to formally address the issue of asymptotic convergence of the HPT process. While this study has numerically established that HPT converges for varying signal types and initial frequency vectors, a complementary mathematical investigation that examined convergence (say, by randomly perturbing a true frequency vector and examining the off-diagonal terms) might be instructive and lead to improvements in the iterative process.

8.3 Engineering Issues with Harmonic Phase Tracking

This section is divided into three subsections focusing on: improvements to the HPT numerical algorithm, comparison between HPT and FFT representations, and engineering applications.

8.3.1 Numerical Implementation of HPT

Chapter 4 and Appendix C describe the HPT technique as implemented here. As stated elsewhere in the text, a conservative approach was used to maximize robustness. Improvements to the general technique are surely possible as demonstrated by the following suggested topics:

1. The rules regarding when adjacent harmonics should or must be merged were kept simple. They are almost certainly not optimum and are probably responsible for increasing the number of iterations in some cases.

2. Similarly, the rules regarding the insertion of pairs of sinusoids was kept simple, and they are also suspected of increasing the number of iterations. In particular, if only one of the new amplitudes has an amplitude larger than the HPT threshold, then only it should be used to minimize the rank of the new basis matrices (indexing in the present coding was simplified by always inserting pairs).

3. The total least squares solutions versus the forward and backward time shifts within each iteration are presently independent. Since the smallest time shifts are presently only one point, it is suggested that a recursive scheme be investigated to minimize computational time.

4. One of the fundamental steps within HPT is adjusting the trial frequencies at each iteration. The present implementation conservatively increases or decreases the frequency subject to a variety of ad-hoc stability constraints that were developed and evaluated as problems arose. This is perhaps the most important step to critically examine in a future study because a more accurate process would greatly improve the convergence.

5. Global controls on the convergence are likewise simple. For example, if HPT is struggling to achieve the desired accuracy in the time domain error by inserting new sinusoids over many successive iterations, then a variety of internal thresholds are concluded to be unrealistic and are increased slightly. In other cases, the present code has been observed to repeatedly delete and insert the same sinusoid pair (when the "true" sinusoids are spaced just lower than the HPT resolution and the internal controls act independently); improved recognition of this phenomenon and subsequent adjustment of the controls to avoid it would improve convergence time.

Generally speaking, the "first generation" HPT implementation used in this study should be critically reexamined, streamlined, and made as consistent as possible.

8.3.2 Comparison Between HPT and FFT Representations

Since FFT-based spectra are the most commonly used for engineering spectral estimates, this section augments the comparison between them and HPT representations present throughout this study by examining four new topics.

The first topic is frequency resolution of HPT estimates. The complication here is the definition of the number of points used by HPT to achieve the stated resolution. On the surface, HPT clearly has a smaller resolution than the FFT when the reference is the HPT segment length. But strictly speaking, HPT requires forward and backward time shifts to identify the frequency vector (although the raw amplitudes and phases for each shift are defined using the original segment length). Thus, from an information theory perspective, it can be alternatively concluded that HPT resolution is comparable to FFT resolution when the total number of data points are used as the reference. Both interpretations are right.

The second topic is windowing. On this topic there is no ambiguity: the use of a window is unnecessary in HPT. Given the tremendous attention

paid in the literature to the properties and unresolved merits of windowed versus non windowed FFTs, eliminating their use has a strong attraction. Windows are, of course, undefined for use with HPT because it does not artificially impose periodicity and the subsequent end effects in the signal.

The third topic is the fact that the segment length for HPT is independent of the periodicity for deterministic signals. As an example, a unit amplitude square wave signal with a period of 48 points was analyzed using a HPT segment length of 60 points. The HPT frequencies were referenced to a 32-point FFT, which is simply the closest multiple-of-2 length relative to the 48 point period. Thus, the first exact harmonic appears at bin number of $^{32}/_{48}$, or 0.667 cycles per the reference FFT length, with significant energy only at additional odd harmonics corresponding to bin numbers of 0.667*[3 5 7 ...] = 2, 3.333, 4.667, 6. etc. In addition, the exact amplitudes at those odd harmonics would be $(^{4}/\pi)$/[1 3 5 7 9 11 ...] (Carslaw, 1930; Dean and Dalrymple, 1984). The HPT-estimated versus exact parameters for this square wave are compared below:

Harmonic	Bin Number		Amplitude	
	Exact	HPT	Exact	HPT
1	0.6667	0.6670	1.2732	1.2756
3	2.0001	1.9996	0.4244	0.4263
5	3.3335	3.3294	0.2546	0.2553
7	4.6669	4.6636	0.1819	0.1909
9	6.0003	5.9917	0.1415	0.1548
11	7.3337	7.3320	0.1157	0.1229
13	8.6671	8.6640	0.0979	0.1072

Table 8.1 Exact HPT Estimated Bin Numbers and Amplitudes for a Square Wave Signal

Again, these accuracies were achieved: (1) without knowing the period of the signal, (2) without applying any window, and (3) without averaging the first data point as is required for FFT analyses. The HPT convergence error threshold was 0.5 percent for the rms value of the time domain error.

This fourth and last discussion compares the utility of the FFT and HPT techniques, where the orthogonality of FFT estimates is shown to offer advantages compared to HPT estimates for some signal descriptors. For example, Parseval's Theorem relating the time versus frequency domain mean square value of a signal is computationally very efficient using the

equally-spaced FFT ordinates. Conversely, the nonorthogonality of the HPT estimates complicates the equivalent frequency domain calculation. The mean square value Ψ^2 over a finite interval is defined as:

$$\psi^2 \equiv \int_{-\xi}^{\xi} x^2(t)\,dt \qquad 8.5$$

The HPT representation for the signal can be written in vector form as:

$$\begin{aligned} \mathbf{x} &= \mathbf{a} \bullet \mathbf{r} + \mathbf{b} \bullet \mathbf{i} \\ &= [\mathbf{r} \;\; \mathbf{i}] \begin{Bmatrix} \mathbf{a} \\ \mathbf{b} \end{Bmatrix} \\ &= \mathbf{t}^T \mathbf{c} \end{aligned} \qquad 8.6$$

where vectors $\mathbf{r} \equiv \cos(2\pi \mathbf{f}\mathbf{t})$ and $\mathbf{i} \equiv \sin(2\pi \mathbf{f}\mathbf{t})$ based on the rank R frequency vector $\mathbf{f}$, $\bullet$ denotes the inner (dot) product operator, and ^{T} is the transpose operator. Substitute the last equation into a vector version of Equation 8.5:

$$\begin{aligned} \psi^2 &\equiv \int_{-\xi}^{\xi} \mathbf{x}^T \mathbf{x}\,dt \\ &= \int_{-\xi}^{\xi} [\mathbf{t}^T \mathbf{c}]^T [\mathbf{t}^T \mathbf{c}]\,dt \\ &= \int_{-\xi}^{\xi} \mathbf{c}^T \mathbf{t}\mathbf{t}^T \mathbf{c}\,dt \\ &= \mathbf{c}^T \left[\int_{-\xi}^{\xi} \mathbf{t}\mathbf{t}^T\,dt\right] \mathbf{c} \end{aligned} \qquad 8.7$$

The last step is valid since $\mathbf{c}$ is not a function of time. Now observe that

$$\int \mathbf{t}\,\mathbf{t}^T\,dt = \int [\mathbf{r}\,\mathbf{i}] \begin{bmatrix} \mathbf{r}^T \\ \mathbf{i}^T \end{bmatrix} dt$$

$$= \begin{bmatrix} \int \mathbf{r}\,\mathbf{r}^T\,dt & \mathbf{0} \\ \mathbf{0} & \int \mathbf{i}\,\mathbf{i}^T\,dt \end{bmatrix} \qquad 8.8$$

After further simplification the mean square expression becomes

$$\psi^2 \equiv \mathbf{c}^T \left[\int \mathbf{t}\,\mathbf{t}^T\,dt \right] \mathbf{c}$$

$$= \mathbf{a}^T \left[\int \mathbf{r}\,\mathbf{r}^T\,dt \right] \mathbf{a} + \mathbf{b}^T \left[\int \mathbf{i}\,\mathbf{i}^T\,dt \right] \mathbf{b} \qquad 8.9$$

The two R-by-R matrices in brackets were previously defined in Equation 4.30 as integral to the HPT methodology. So, while HPT does require two full matrix multiplications because the frequencies are not orthogonal, at least the matrices would already be resident from the HPT analysis.

Another interesting aspect of signal processing where Fourier Series can be more informative than HPT is the identification of frequency response functions. Fitting a single sinusoid with a non integer period using Fourier Series yields amplitudes at all frequencies; if an excitation and response are both fitted, then cross spectra can be used to estimate the system operator over a wide frequency range. Conversely, HPT would [correctly] estimate energy only at one discrete frequency, and would accordingly only estimate the system operator at that one frequency.

8.3.3 Engineering Applications of HPT

The real significance of HPT is its ability to identify physically meaningful harmonic components from measured data. For many tasks such as those discussed in the previous section, the orthogonality of the Fourier Series offers important and useful information. The objective of this section is to illustrate where HPT information can enhance engineering and scientific knowledge compared to the use of Fourier Series.

HPT has the potential for advances in signal extrapolation. Figure 8.1a illustrates HPT wave extrapolation using waves near the peak of Hurricane Bob. This figure was previously introduced and explained as Figure 5.7. This example demonstrates a high level of correlation between the extrapolated and measured signal. This correlation is a function of the stationarity of the signal components along with the starting index of the chosen segment to be extrapolated, and the sophistication of the HPT fitting technique. For example, Figure 8.1a extrapolates relative to the center segment; in other words, it is "extrapolating" into a section of the signal that was already used indirectly for the forward and backward time shifts in the frequency fitting. But this extrapolation was done using the amplitudes and phases fitted only to the center segment (time indices between 0 and 300 in the Figure). Figure 8.1b illustrates extrapolation for the same signal but into a segment of the time series not used in the HPT fit

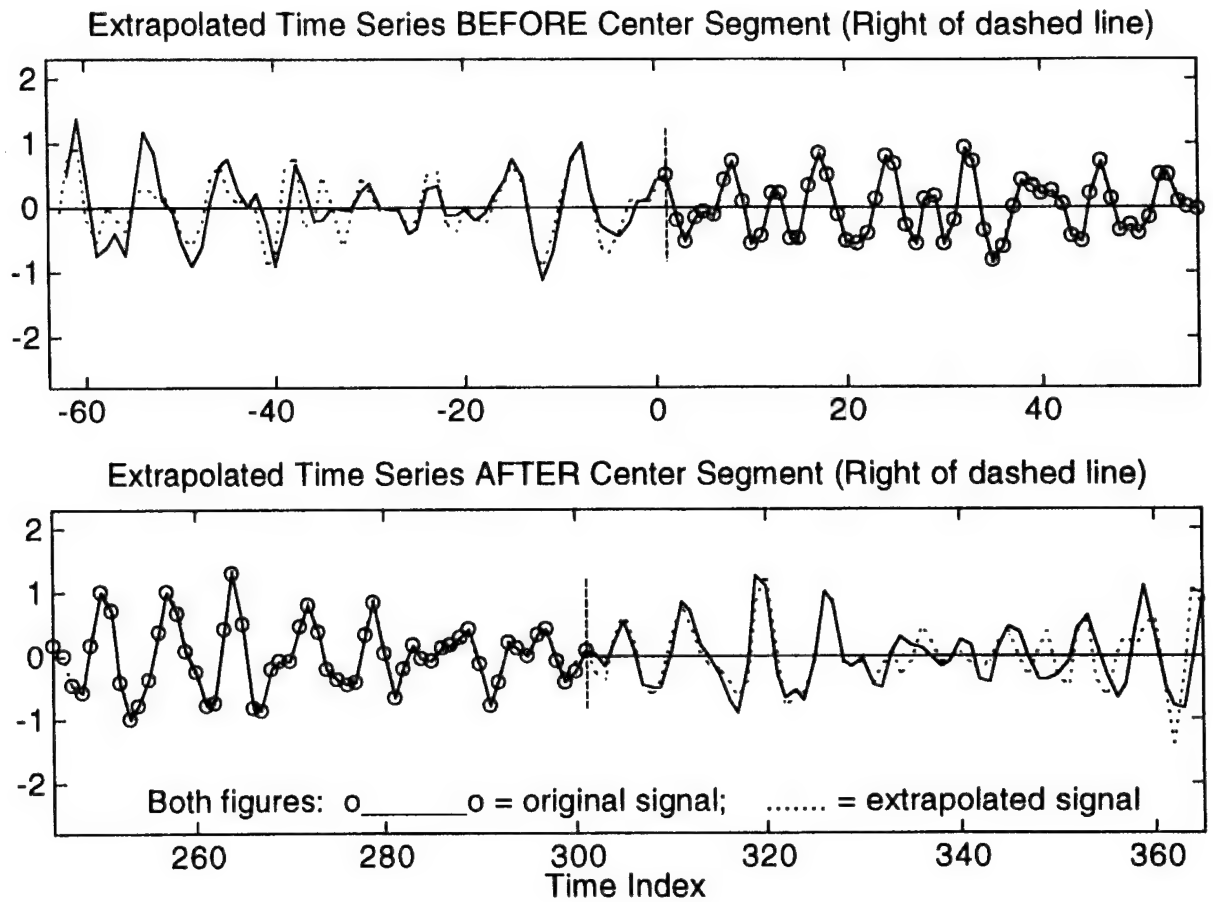

Figure 8.1a Representative HPT Signal Extrapolation Relative to Center Segment Using Waves from Hurricane Bob.

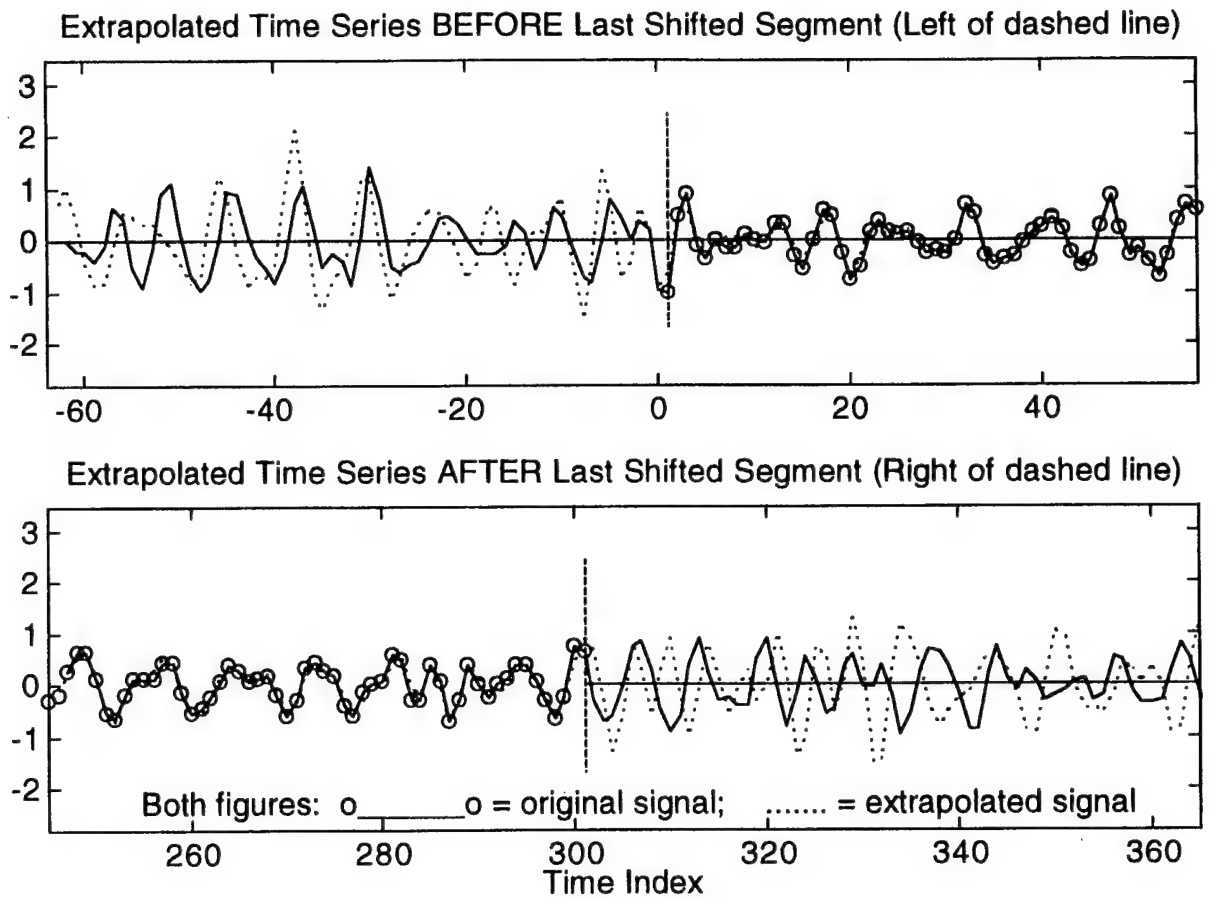

Figure 8.1b Representative HPT Signal Extrapolation Relative to Stationary Frequency Vector Using Waves from Hurricane Bob.

(indices above 300 in the lower figure). This fit assumes stationarity only of the frequency vector to perform a least squares fit for new amplitudes and phases relative to the most forward time shifted segment used for the HPT analysis, then extrapolates using those parameters. This fit is not as well correlated as before, which in this case reflects nonstationarity in the wave field. There are improvements that can be made to this simple process; for example, a series of HPT raw estimates could be averaged to find slowly-varying "averaged" behavior of the frequency vector, which could then be incorporated into the extrapolation model. Second, the process could be made more computationally attractive by investigating recursive HPT methods, and by truncating the rank of the HPT estimates to only include significant components. The engineering potential of HPT for projecting future events, such as several cycles of an incident wave for ship dynamics, is a prime topic for investigation.

HPT also provides several opportunities to advance statistical measures of stationarity for any signal. For example, rms error of extrapolations may serve as a useful scalar-type measure. The phase continuity figures used extensively in Chapter 7 may provide a very qualitative vector measure of stationarity in terms of component continuity. Finally, variations in frequency and amplitude evolutions are a direct vector measure of stationarity (similar to "order N" descriptors presently used). Such descriptors may provide a more rigorous framework for evaluating stationarity compared to existing statistical and probabilistic descriptors.

By identifying true signal frequencies, HPT may also contribute to linear and nonlinear system studies. The conceptual justification for this is straightforward to illustrate. For example, the response of a single sinusoid passing through a linear system must be another (shifted and scaled) single sinusoid at the same frequency. Yet, if the frequency is in-between Fourier Series harmonics, then a FFT representation of that single sinusoid would have amplitudes at all frequencies up to the Nyquist. Passing that "spread" transform through a linear system, particularly a lightly damped system with a resonant frequency not equal to the excitation frequency, would alter the amplitudes and phases such that the response would not be a single sinusoid. Of course, windowing the signals limits this spread in the transforms, but the principle is the same. For example, the wave pressures at the FRF array are transformed then passed through a scale factor (filter) to recover the surface wave amplitudes; a HPT analysis of this process might yield a different surface wave profile. Also, for nonlinear systems, the frequency information from HPT might be used to better isolate N^{th}-order forced super- or sub-harmonics from neighboring free components or noise.

With regards to identifying true signal properties, HPT has applications in almost every field of engineering and science. Speech recognition is a good candidate for a HPT application. Other ocean engineering applications are possible such as in sonar signal processing, where the

nonstationarity of contacts and the Doppler shifts are both difficult to identify with orthogonal techniques.

This study has emphasized engineering applications using laboratory and geophysical water waves. It was beyond the scope of this one study to definitively quantify the behavior of an ocean wave field. But this study has tried to demonstrate that HPT provides new tools that in many cases for the first time directly address phenomena associated with such fields, such as:

- continuity of components (wave group run lengths),
- stationarity of the wave field (evolutionary plots)
- downshifting (evolutionary plots)
- shortcrestedness (spatial coherence)
- incident wave direction and wavelengths

Other information from further studies are also recommended to resolve some intriguing observations from this preliminary wave study. Figure 8.2 shows the evolution of waves near the spectral peak for the stationary wave field used in Chapter 7 on September 13 1991. 10 minute HPT analyses were used. In general, the representative packets that have had lines superimposed show consistent downshifting at rates which seem large for a stationary field. Second, these packets apparently persist over very long time intervals. And third, there are at least two instances shown in Figure 8.2 where two packets merge (apparent because the new amplitude is

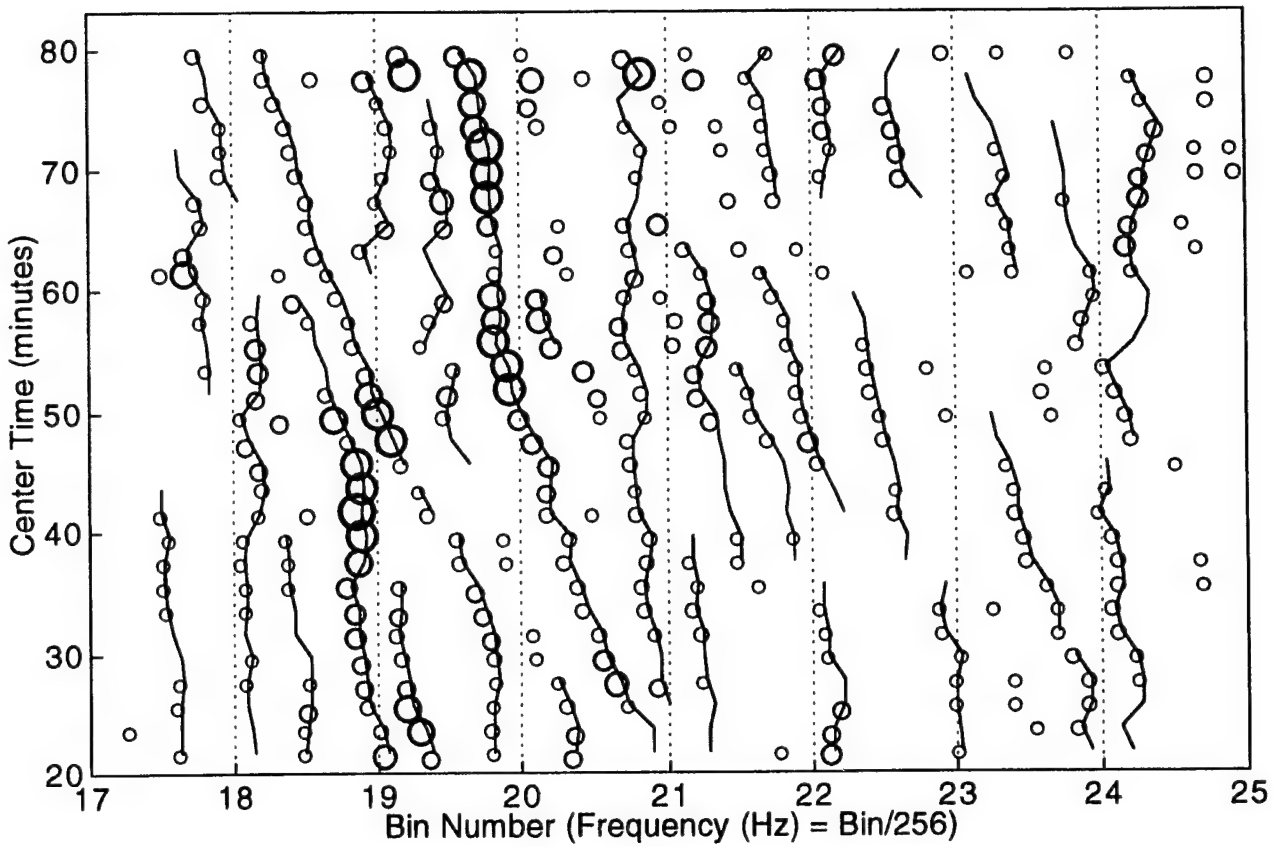

Figure 8.2 Wave Packet Evolution, September 13 1991, Gage 211.

relatively large, implying that it is a resultant of the energy in the previous two packets). Whether, how, or why these packets are merging is an open question.

This Chapter on a perspective of HPT ends with a quote (Dyson, 1995);

> *The great advances in science usually result from new tools rather than from new doctrines. ..... Science flourishes best when it uses freely all the tools at hand, unconstrained by preconceived notions of what science ought to be. Every time we introduce a new tool, it always leads to new and unexpected discoveries, because Nature's imagination is richer than ours.*

It is hoped that Harmonic Phase Tracking becomes a useful new tool that will lead to many new discoveries.

REFERENCES

Apel, J. R., Principles of Ocean Physics, Academic Press, 1987.

Athanassoulis, G.A., Vranas, P. B., and Soukissian, T. H., A New Model for Long-Term Stochastic Analysis and Prediction- Part I: Theoretical Background, Journal of Ship Research, v36, n1, Mar 1992, pp 1-16.

Bendat, J. S., and Piersol, A. G., Random Data Analysis and Measurement Procedures, 2nd Ed, John Wiley and Sons, 1986.

Bendat, J. S., and Piersol, A. G., Engineering Applications of Correlation and Spectral Analysis (2nd Ed), John Wiley and Sons, 1993.

Berkemeier, W. A., Miller, H. C., Wilhelm, S. D., DeWall, A. E., and Gorbics, C. S., A User's Guide to the Coastal Engineering Research Center's (CERC's) Field Research Facility, Instruction Report CERC-85-1, U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS, May, 1985.

Billingsley, P., Probability and Measure, 2nd Ed, John Wiley and Sons, 1986.

Bitner-Gregersen, E. M., and Gran, S., Local properties of sea waves derived from a wave record, Applied Ocean Research, v5 n4, 1983, pp210-214.

Boashash, B., Time-Frequency Signal Analysis, Longman Cheshire, 1992.

Borgman, L. E., Petrakos, M., and Li, C., Evolutionary Fourier Analysis of Wave Data, Ocean Wave Measurement and Analysis, ASCE, 1994.

Carslaw, H. S., An Introduction to the Theory of Fourier's Series and Integrals, 3rd Ed, Dover, 1930.

Chu, P., The S-transform for Obtaining Localized Spectra, Marine Technology Society Journal, v29 n4, 1996, pp 28-38.

Dean, R. G., and Dalrymple, R. A., Water Wave Mechanics for Engineers and Scientists, Prentice-Hall, 1984.

Defant, A., Physical Oceanography, Pergamon Press, 1961.

Donelan, M. A., and Drennan, W. M., Nonstationary Analysis of the Directional Properties of Propagating Waves, Journal of Physical Oceanography, v26, Sept 1996, pp1901-1914.

Dyson, F., The Scientist As Rebel, in Nature's Imagination, The Frontiers of Scientific Vision, J. Cornwell (Ed), Oxford University Press, 1995.

Cornwell, J. (Ed), Nature's Imagination. The Frontiers of Scientific Vision, Oxford University Press, 1995.

Elgar, S. L., and Seymour, R. J., Effects of the Lack of Stationarity on Deep Water Wave Statistics, Oceans 85 Conference, Nov 1985, San Diego, CA, pp 718-722.

Field Research Facility, Preliminary Data Summary September 1990, U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS.

Field Research Facility, Preliminary Data Summary August 1991, U.S. Army Corps of Engineers, Waterways Experiment Station, Vicksburg, MS.

Goda, Y., Random Seas and Design of Maritime Structures, University of Tokyo Press, 1985.

Huang, N., Long, S., Chi-Chao, T., Donelan, M, Yuan, Y. and Lai, R., The Local Properties of Ocean Surface Waves by the Phase-Time Method, Geophysical Research Letters, v19, Apr 3 1992, pp685-688.

Hughes, S., Spatial Variability in the Nearshore Wavefield, U. S. Army Waterways Experiment Station Miscellaneous Paper CERC-84-7, Vicksburg, MS, July 1984.

Horikawa, K., Nearshore Dynamics and Coastal Processes, University of Tokyo Press, 1988.

Jammalamadaka, S. R., and Sama, Y. R., Circular Regression, in Statistical Sciences and Data Analysis, Eds K. Matusita, et. al., VSP, Utretch Netherlands, 1993, pp109-128.

Jenkins, G. M., and Watts, D. G., Spectral Analysis and its Applications, Holden-Day, 1968.

Kay, S. M., Modern Spectral Estimation, Prentice-Hall, 1988.

Kay, S. M., and Marple, S., L., Jr., Spectrum Analysis - A Modern Perspective, IEEE Proceedings, v69, n11, Nov 1981, pp1380-1419.

Kinsman, B., Water Waves, Dover, 1984.

Komen, G. J., Cavaleri, L, Donelan, M., Hasselmann, K., Hasselmann, S., Jannssen, P. A. E. M., Dynamics and Modelling of Ocean Waves, Cambridge University Press, 1994.

Leon, S. J., Linear Algebra with Applications, Macmillan, 4th Ed, 1994.

MATLAB Users Manual, The Math Works, South Natick MA, 1989.

Medina, J. R., and Husdpeth, R. T., A Review of the Analyses of Ocean Wave Groups, Coastal Engineering, 14 (1990), pp 515-542.

Montgomery D., and Peck E., Introduction to Linear Regression Analysis, Wiley, 1992.

Naidu, P. S., Modern Spectrum Analysis of Time Series, CRC Press, 1996.

Papoulis, A., Probability, Random Variables, and Stochastic Processes, McGraw-Hill, 1965.

Roy III, R. H., ESPRIT: Estimation of Signal Parameters via Rotational Invariance Techniques, PhD Thesis, Stanford University, 1987.

Schmidt, R. O., A Signal Subspace Approach to Multiple Emitter Location and Spectral Estimation, PhD dissertation, Stanford University, May 1979.
Sorensen, R. M., Basic Wave Mechanics for Coastal and Ocean Engineers, John Wiley and Sons, 1993.

Trevino, G., The Frequency Spectrum of Nonstationary Random Processes, in Time Series Analysis: Theory and Practice 2 (Anderson, O. D., Editor), North-Holland , 1982, pp 237-247.

Toba, Y., Okada, K., and Jones, I. S., The Response of Wind-Wave Spectra to Changing Winds. Part I: Increasing Winds, Journal of Physical Oceanography, v18 n9 (1988), pp1231-1240.

Tucker, M. J., Nearshore Waveheight During Storms: Reply to the Comments of S. A. Hughes, Coastal Engineering, 26 (1995), pp109-115.

Van Huffel, S., and Vandewalle, J., The Total Least Squares Problem - Computational Aspects and Analysis, SIAM, 1991.

Wang, W., Wavelet Transform in Vibration Analysis for Mechanical Fault Diagnosis, Shock and Vibration, v3 n1, 1996, pp17-26.

Werle, B. O., Sea Backscatter, Spikes and Wave Group Observations at Low Grazing Angles, IEEE International Radar Conference, 1996.

Zseleczky, J. and Wallendorf, L., Extreme Roll Motions of a Navy Frigate Model in Large Beam Seas, Hydromechanics Laboratory, United States Naval Academy, Annapolis, MD, Oct 1994.

APPENDIX A

USEFUL ALGEBRA FOR MULTIHARMONIC SIGNALS

One basic building block of the new Harmonic Phase Tracking parameter estimation technique is paired sinusoids, which are used to model multiharmonic signals when the frequencies are close to the frequency resolution. At first glance it seems unnecessary to review the fundamental behavior of a model as simple as paired sinusoids. But looks can be deceiving. For example, if asked to describe the sum of two sinusoids with constant parameters, a large number of engineers and scientists would dismiss it as a trivial beating behavior. However, closer algebraic evaluation yields a surprising behavior that requires more careful handling. Knowledge of this behavior also allows for a more accurate assessment of real-world signals with narrowband and/or a continuous spectrum, and an appreciation of how techniques such as Fourier Series and Harmonic Phase Tracking model such signals. Understanding beating behavior and its consequences is therefore an important prerequisite for correct modeling of ocean waves.

A.1 General Algebraic Expressions for Two Summed Sinusoids.

As stated above this is a basic building block for this technique, as well as the clearest model for demonstrating the algebra of multiharmonic signals. Define a deterministic signal comprised of two sinusoids, with unit amplitudes and arbitrary frequencies and phases:

$$x(t) = \sum_{i=1}^{2} \cos\left(2\pi f_i t + \theta_i\right) \qquad \text{A.1}$$

The fact that both amplitudes are equal ($a_1 = a_2 = 1$) is important for this first example. There are several alternative algebraic expressions to Equation A.1 that better describe the resulting signal. It will be useful for most of the following expressions to define a mean frequency and a [half] difference frequency:

$$\bar{f} = \frac{f_2 + f_1}{2} \qquad \text{A.2a}$$

$$f_\Delta = \frac{\left|f_2 - f_1\right|}{2} \qquad \text{A.2b}$$

Equation A.2 is a direct result of the familiar set of trigonometric identities for the sum of two *unit amplitude* sinusoids: for example,

$$\sin\alpha + \sin\beta = 2\sin\left(\frac{\alpha+\beta}{2}\right)\cos\left(\frac{\alpha-\beta}{2}\right) \qquad \text{A.2c}$$

Also note that the difference frequency in Equation A.2b is more accurately labeled a "half-difference" frequency because it measures the

difference from each frequency to the mean frequency, not between the two frequencies.

The first equivalent expression to Equation A.1 separates x(t) into two products based on in- and out-of-phase components of the mean frequency:

$$x^{(1)}(t) = E_c(t|f_\Delta)\cos(2\pi\bar{f}t) + E_s(t|f_\Delta)\sin(2\pi\bar{f}t) \qquad \text{A.3}$$

where E_c and E_s are time-varying envelopes for the cosine and sine terms, respectively, and the dependence on f_Δ is explicitly shown. Equation A.3 was not further reduced to one trigonometric term proportional to the mean frequency for a reason to be explained shortly. The functional form of Equation A.3 as shown is important. Note that the mean frequency trigonometric terms describe the "instantaneous" signal that would be most apparent upon inspection of the total signal. These terms are in turn multiplied by "slowly-varying" or "modulating" envelopes E which are functions of time and the difference frequency.

The various terms are now expanded for completeness. The time-varying amplitude (envelope) for the cosine term is denoted by $E_c(t|f_\Delta)$ and is specified by:

$$E_c(t|f_\Delta) \equiv A_c\cos(2\pi f_\Delta t + \varphi_c) \qquad \text{A.4a}$$

while the similar envelope term for the out-of-phase sine term is

$$E_s(t|f_\Delta) \equiv A_s\sin(2\pi f_\Delta t + \varphi_s) \qquad \text{A.4b}$$

A_C and A_S are positive constants defined by:

$$A_c = \sqrt{(a_1 \sin\theta_1 + a_2 \sin\theta_2)^2 + (a_2 \cos\theta_2 - a_1 \cos\theta_1)^2} \qquad \text{A.5a}$$

$$A_s = \sqrt{(a_1 \cos\theta_1 + a_2 \cos\theta_2)^2 + (a_1 \sin\theta_1 - a_2 \sin\theta_2)^2} \qquad \text{A.5b}$$

The two amplitudes a_1 and a_2 are still defined as equal but are included for completeness, while the respective phases (subject, as all the phases in Appendix A are, to proper quadrant placement such as with the ATAN2 function in MATLAB) are given by:

$$\varphi_c \equiv \tan^{-1}\left\{\frac{a_2 \cos\theta_2 - a_1 \cos\theta_1}{a_1 \sin\theta_1 + a_2 \sin\theta_2}\right\} \qquad \text{A.6a}$$

$$\varphi_s \equiv \tan^{-1}\left\{\frac{a_1 \sin\theta_1 - a_2 \sin\theta_2}{a_1 \cos\theta_1 + a_2 \cos\theta_2}\right\} \qquad \text{A.6b}$$

Equation A.3 is simplified one last time to produce

$$x^{(1)}(t) = \sqrt{E_c(t|f_\Delta)^2 + E_s(t|f_\Delta)^2}\cos(2\pi\bar{f}t + \tan^{-1}\left\{\frac{E_s(t|f_\Delta)}{E_c(t|f_\Delta)}\right\}) \qquad \text{A.7}$$

Although the behavior of the time-varying envelope term is not readily apparent in this formulation, it is clearly seen that it is always a positive (not a zero mean) function. Note also the time-varying phase.

Return to Equation A.3 which clearly shows that both of the trigonometric functions of the mean frequency are modulated by envelopes that are relatively slowly-varying. However, it also shows a second more subtle behavior, described below, that creates tremendous complications in

modeling and interpreting the total signal. Recall that the envelope given in Equation A.7 is always positive, which follows convention for describing modulating envelopes as non-negative. However, the component envelopes defined in Equation A.3 are zero-mean trigonometric functions, meaning that they take on negative values.

Something unexpected happens to either component (product) term in Equation A.3 when the modulating envelope changes sign. If that one component is examined graphically or algebraically, the "instantaneous" component at the mean frequency will show an abrupt sign reversal, in other words, an *instantaneous 180 degree phase shift* where the envelope has a node and the derivative of the rectified envelope is infinite. This phenomenon is illustrated in Figure A.1. Note how the phase of the sum-frequency sinusoid and the effective, beating sinusoid go from in-phase to out-of-phase across the node of the envelope.

Now, note the similar behavior of representative ocean waves from Hurricane Bob shown in Figure A.2, confirming that this 180 degree phase discontinuity does occur in real-world signals and is not simply a mathematical curiosity using idealized sinusoids. (Demonstration of the existence of this phase discontinuity phenomenon for real ocean waves also shows that zero-crossing methods can be lead to biased estimates.)

This phase discontinuity is further explored in subsequent sections.

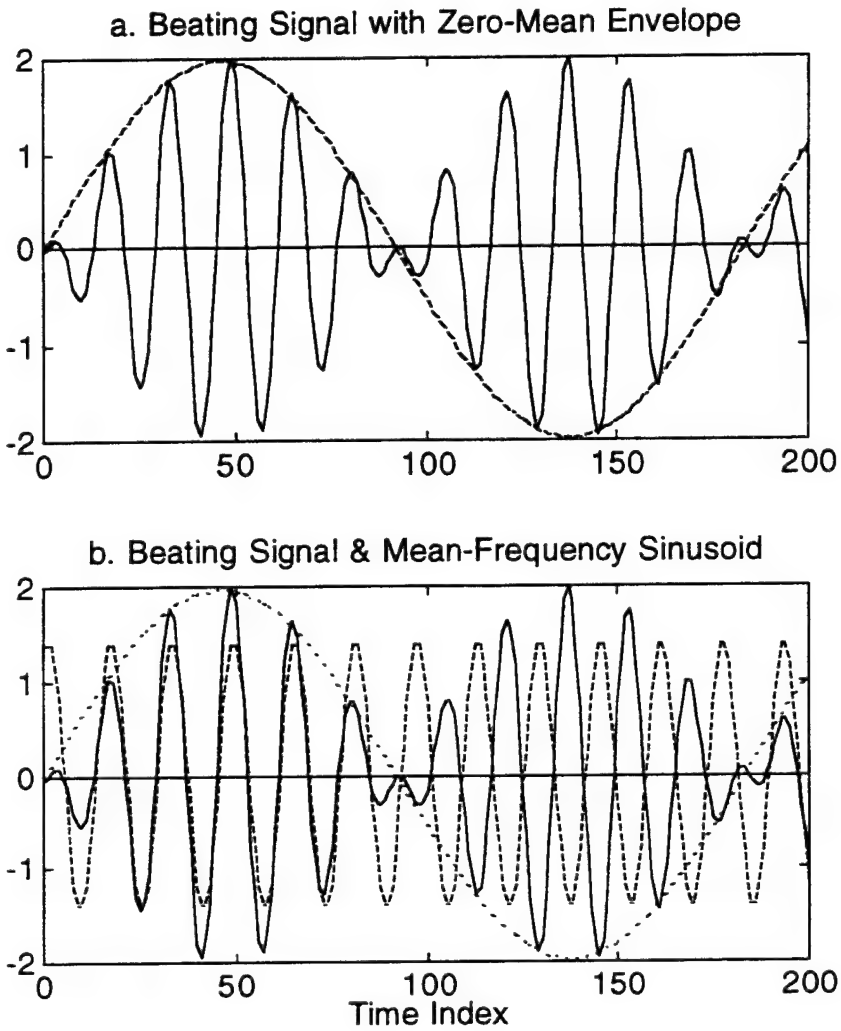

Figure A.1 Illustration of Phase Discontinuity in Beating Sinusoids. Figure A.1.a: instantaneous signal with zero-mean envelope; Figure A.1.b: same signal with sum-frequency sinusoidal signal (- - - -).

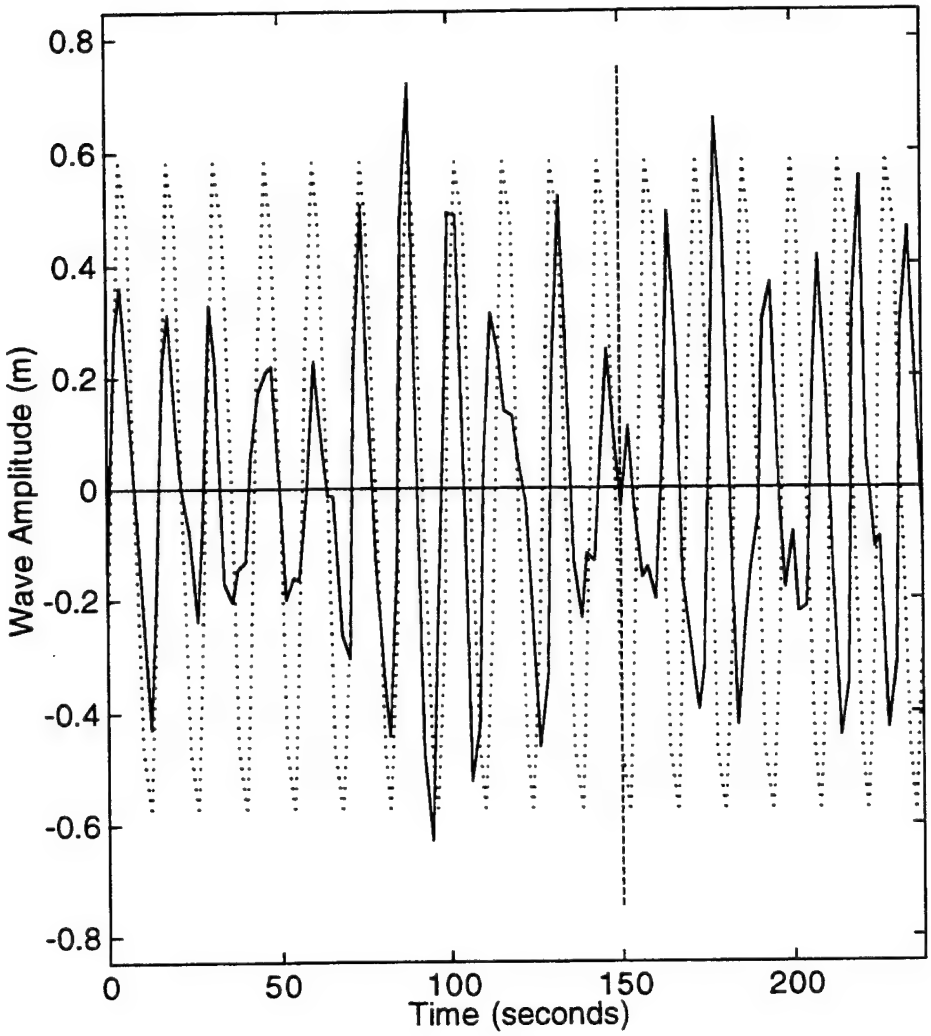

Figure A.2 Example of Phase Discontinuity (at 150 sec) in an Ocean Wave Signal (Prior to Hurricane Bob, Gage 131, at 1945 on Aug 18 1991); ___ = wave signal; . . . . = constant-parameter sinusoid

Equations A.3 and A.7 completely describe the resulting signal for equal amplitudes and any values of frequency and phase. But not all combinations are worthy of attention. Consider the frequency separation; if the two frequencies are not close, then the signal looks like a fast sinusoid riding on a slow sinusoid. In the complimentary limit as the two frequencies come together, the difference frequency approaches zero so the envelope period tends towards infinity, and over a finite interval the signal appears as one sinusoid with an quasi-stationary amplitude (except right at a node). For these situations where the difference period is much longer than the length of the finite segment being analyzed, the signal can appear to be a constant frequency sinusoid with a nonstationary amplitude.

The "simplicity" of the previous beating signal is greatly complicated when the amplitudes are unequal. At the limiting case previously described when both amplitudes are equal, the classic beating results, with a symmetrical envelope that causes a 180 degree phase shift in the instantaneous signal. At the other limit where one of the amplitudes becomes negligible, an apparent single sinusoid results. But for the intermediate case when one of the amplitudes is a finite fraction of the other, the signal modulates about a "mean amplitude" and the envelope never approaches a zero value. From visual inspection of such an "intermediate" beating signal one might conclude that the sign reversal phenomenon never occurs. But Equations A.9 and A.10 (below) shows that

it always occurs somewhere for both the in-phase and out-of-phase instantaneous components, and at any arbitrary phase (i.e., time shift) relative to the other. In other words, this sign reversal may occur when the envelope is at a "node", or it may occur at some other time. As described in the next section on Instantaneous Frequency, this arbitrary reversing is a serious problem for "real world" signals.

This all leads to two additional alternative algebraic expressions for two beating sinusoids with arbitrary amplitudes that are generally applicable and analytically informative. The first starts with definitions of a mean amplitude and a difference amplitude:

$$\overline{A} \equiv \frac{a_1 + a_2}{2} \qquad \text{A.8a}$$

$$A_\Delta \equiv \frac{|a_1 - a_2|}{2} \qquad \text{A.8b}$$

such that $a_1 \equiv \overline{A} + A_\Delta$ and $a_2 \equiv \overline{A} - A_\Delta$. Similarly, define mean and difference phase functions:

$$\overline{\theta} \equiv \frac{\theta_1 + \theta_2}{2} \qquad \text{A.8c}$$

$$\theta_\Delta \equiv \frac{|\theta_1 - \theta_2|}{2} \qquad \text{A.8d}$$

Then algebra yields a second representation for this paired signal:

$$\begin{aligned} x^{(2)}(t) = &\left[2\overline{A}\cos(2\pi f^- t + \theta_\Delta)\right]\cos(2\pi f^+ t + \overline{\theta}) + \\ &\left[2A_\Delta \sin(2\pi f^- t + \theta_\Delta)\right]\sin(2\pi f^+ t + \overline{\theta}) \end{aligned} \qquad \text{A.9}$$

where the brackets indicate modulating envelope functions. Also note that Equation A.2 does not apply for the "sum" and "difference" frequencies, so new variables f^+ and f^- are introduced (f^+ is discussed in the next section). Equation A.9 is seen to be a sum of two beating sinusoids, each constructed from sinusoids with equal amplitudes. Each product term shows the zero-mean envelopes responsible for the abrupt sign changes in the individual components. But since the total signal is the sum of the two terms, the instantaneous signal never shows a full 180 degree phase shift. Last, note how easily Equation A.9 simplifies to a single modulating envelope and instantaneous sinusoid when the two amplitudes are equal and $A_\Delta=0$.

The common theme to these algebraic manipulations has been to convert the original sum of two sinusoids with *unequal* amplitudes to a form with *equal* amplitudes. This can also be accomplished directly by choosing either amplitude as a "reference" amplitude, and decomposing the other amplitude into two parts. This yields a final alternative form:

$$x^{(3)}(t) = a_1\left[\cos\left(2\pi f_1 t+\theta_1\right)+\cos\left(2\pi f_2 t+\theta_2\right)\right]+\left(a_2-a_1\right)\cos\left(2\pi f_2 t+\theta_2\right) \quad \text{A.10}$$

where a_1 was arbitrarily chosen as the reference amplitude. Again, the first term will exhibit a phase discontinuity at some time, but the effective phase discontinuity in the total signal will be lessened (and harder to observe) by the presence of the second [continuous] sinusoid.

A.2 Instantaneous Frequency.

The f^+ sum frequency in Equation A.9 is a function of the two amplitudes and is often called the Instantaneous Frequency (IF). The IF is common in the communications field. Its basis can be understood by consider the problem of determining the frequency of a single sinusoidal signal with constant but unknown parameters. The time-varying trigonometric argument is given by $(2\pi ft+\phi)$. If this trigonometric argument was somehow analytically or numerically known over a finite segment of time, then one way to estimate the frequency would be to take the derivative of this argument with respect to time:

$$\frac{d}{dt}\left[(2\pi ft+\phi)\right]=2\pi f \qquad \text{A.11}$$

from which

$$f=\left(\frac{1}{2\pi}\right)\frac{d}{dt}\left[(2\pi ft+\phi)\right] \qquad \text{A.12}$$

The objective then is to find a way to estimate the vector argument of the sinusoid. The Hilbert Transform ($\mathcal{H}$) is utilized to do this. It can be shown that for positive frequencies the Hilbert Transform of a monochromatic signal is -j*sign(f) times the Fourier Transform where j=sqrt(-1) (Bendat and Piersol, 1986). Thus, it is identically a -90 degree phase shift of the signal, e.g., it shifts a cosine to a sine function and visa versa. This solves the problem as follows; define a Hilbert-transformed signal as:

$$\tilde{x}(t)\equiv\mathcal{H}\{x(t)\} \qquad \text{A.13}$$

Because of the phase shift property, the needed argument function is readily calculated to be:

$$\Omega(t) = (2\pi f t + \phi) = \tan^{-1}\left\{\frac{\tilde{x}(t)}{x(t)}\right\} \qquad \text{A.14}$$

and the instantaneous frequency follows directly from Equation A.12. Of course for a monochromatic signal this numerically-estimated frequency should be constant.

Since the derivative of Equation A.11 can be approximated for a discretely sampled signal using numerical techniques, these expressions are valid at each time step and they can therefore be used to find a time-varying as well as a constant frequency; hence the name "instantaneous frequency". It is mentioned here that a related use for the complex "analytic signal", defined as $x(t) + j\tilde{x}(t)$, is that it can be used to estimate the envelope of a signal (e.g., for a constant sinusoid, $\cos^2\alpha + [\mathcal{H}\{\cos\alpha\}]^2 = \cos^2\alpha + \sin^2\alpha = 1$ as required).

The problem is that the concept of an instantaneous frequency is undefined when more than one sinusoid and/or nonstationary amplitudes are present (Boashash, 1992). However, the lure is that, since the analytic signal is admittedly useful even in these cases for estimating the envelope, it "should" be equally useful for finding the instantaneous frequency. This is not the case however.

To better understand this, consider application of the analytic signal and instantaneous frequency for two [beating] sinusoids. The first obvious reason it has trouble is in handling the abrupt 180 degree phase discontinuity already discussed; a numerically calculated instantaneous phase will detect this discontinuity in the derivative and return essentially a "finite delta function" for the instantaneous frequency at that point. If not anticipated and handled correctly, such an extreme value will greatly skew any short-term, unweighted smoothing of the IF that is routinely done to reduce the effects of noise (additive noise will not be formally incorporated into this discussion as not pertinent to the main point).

Secondly, an analytical expression for the instantaneous frequency of an arbitrary rank summation of sinusoids has been derived (see Boashash, 1992 for further information regarding this review of the IF). The simplified form *for constant parameters* is

$$IF(t) = \frac{\sum_{i=1}^{r} a_i^2 f_i + \sum_{i,j=1\,[i\neq j]}^{r} \frac{1}{2} q_{i,j}(t)\left(f_i + f_j\right)}{\sum_{i=1}^{r} a_i^2 + \sum_{i,j=1\,[i\neq j]}^{r} \frac{1}{2} q_{i,j}(t)} \qquad \text{A.15a}$$

where

$$q_{i,j}(t) = a_i a_j \cos\left(\left[2\pi f_i t + \theta_i\right] + \left[2\pi f_j t + \theta_j\right]\right) \qquad \text{A.15b}$$

Equation A.15a shows two intriguing features:

- IF is a nonlinear function of amplitude, and
- IF is a time-dependent function due to the cross term $q_{i,j}$

Both of these are troublesome. But why is IF a time dependent function when Equation A.9 seems to show the "instantaneous" signal has a constant value? (Note, however, that Equation A.7 confirms the IF time dependence via the time-dependent phase.) The main reason is that the instantaneous frequency is also necessarily a function of the envelope (i.e., varying amplitude) of the signal, because variations in the envelope stretch the signal and in essence stretch the apparent phase. Unfortunately, there is no analytical expression for the IF when the component amplitudes or frequencies are nonstationary.

Consider two additional points. First, the "average frequency" over a finite time segment is formally defined as the average of the IF weighted by the corresponding envelope-squared (this keeps the analog IF finite at the envelope nodes, but not necessarily a numerically-estimated discrete IF). So, the phase discontinuity can still bias an estimate of the "effective, mean" frequency over a finite segment of the signal. Secondly, consider the concept of the time dependent "frequency bandwidth" relative to the IF; this function quantifies the spread of frequencies contributing to the IF at any given time (i.e., it is zero for a single constant frequency sinusoidal signal). The analytical expression for this bandwidth is a function of the first and second time derivatives of the (zero mean) envelope - reinforcing how the envelope affects the "local" frequency.

All of these phenomena, particularly the large effect that the changing amplitude has on the instantaneous frequency, are readily apparent in numerical studies. Figure A.3 illustrates the instantaneous frequency for two signals comprised of two constant-parameter sinusoids. The original signal with two unit amplitudes is shown in the upper figure. The middle figure shows the two component frequencies and the numerically-derived instantaneous frequency; the negative delta functions are caused by the 180 phase discontinuity in the original signal. The lower figure shows the constant and instantaneous frequencies when $a_2 = a_1/2$ (signal not shown). As previously discussed, this softens the effect of the phase discontinuity. But equally as important, this lower figure shows how far the instantaneous frequency can vary away from the two component frequencies (dashed lines).

One additional figure is presented to reinforce this latter observation about the variability and undependability of the instantaneous frequency. Figure A.4 is identical to Figure A.3 except that a third unit amplitude component sinusoid has been added to the two cosinusoids used for Figure A.3. Both the middle and lower subfigures show similar behavior to Figure A.3. Note that the equal amplitudes result in an instantaneous frequency for the middle subfigure which is symmetrical about the three true frequencies; this is expected because the instantaneous frequency is symmetrically weighted by the amplitudes. In the lower subfigure, however, the unequal amplitudes expectedly biased the estimate towards

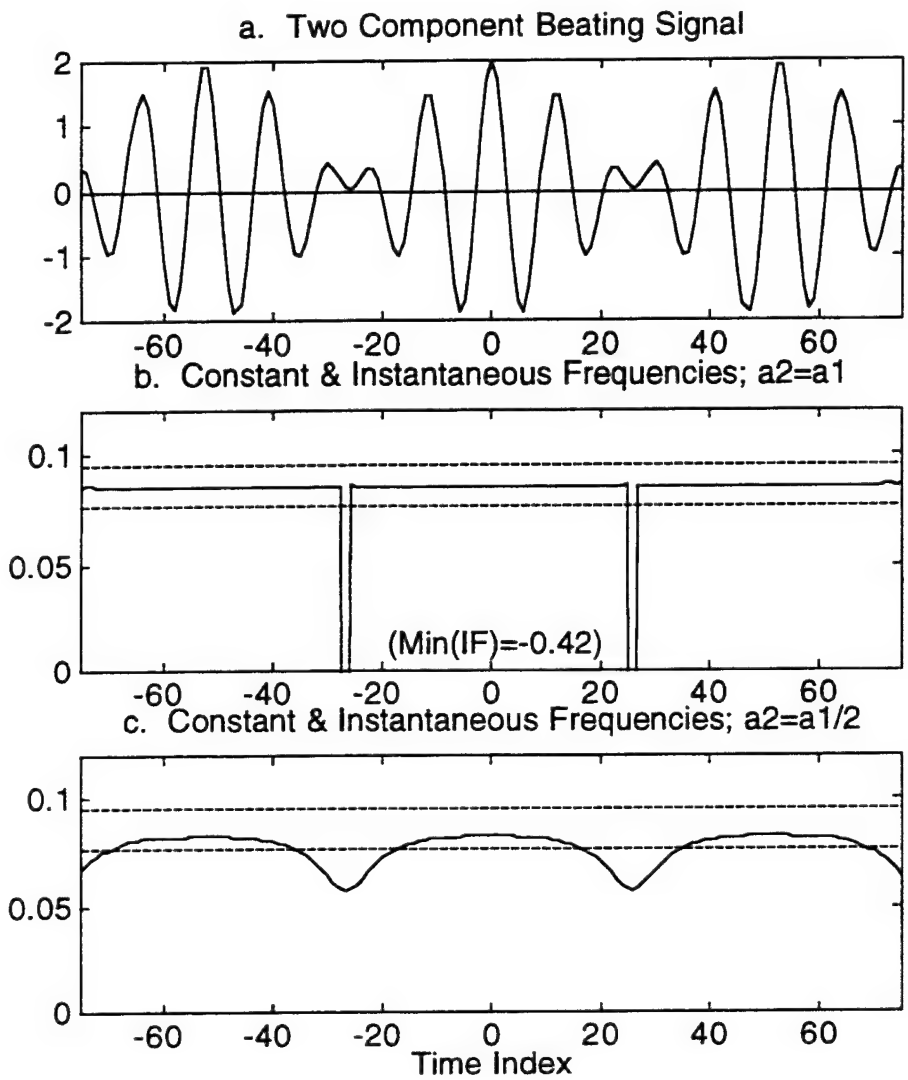

Figure A.3 Illustration of Instantaneous Frequency for Two Sinusoids; a: original signal with equal amplitudes; b: constant ($^{12.2}/_{128}$ and $^{9.75}/_{128}$; dashed line) and instantaneous frequency (solid line) when $a_2=a_1$; c: constant and instantaneous frequency when $a_2={}^{a_1}/_{2}$.

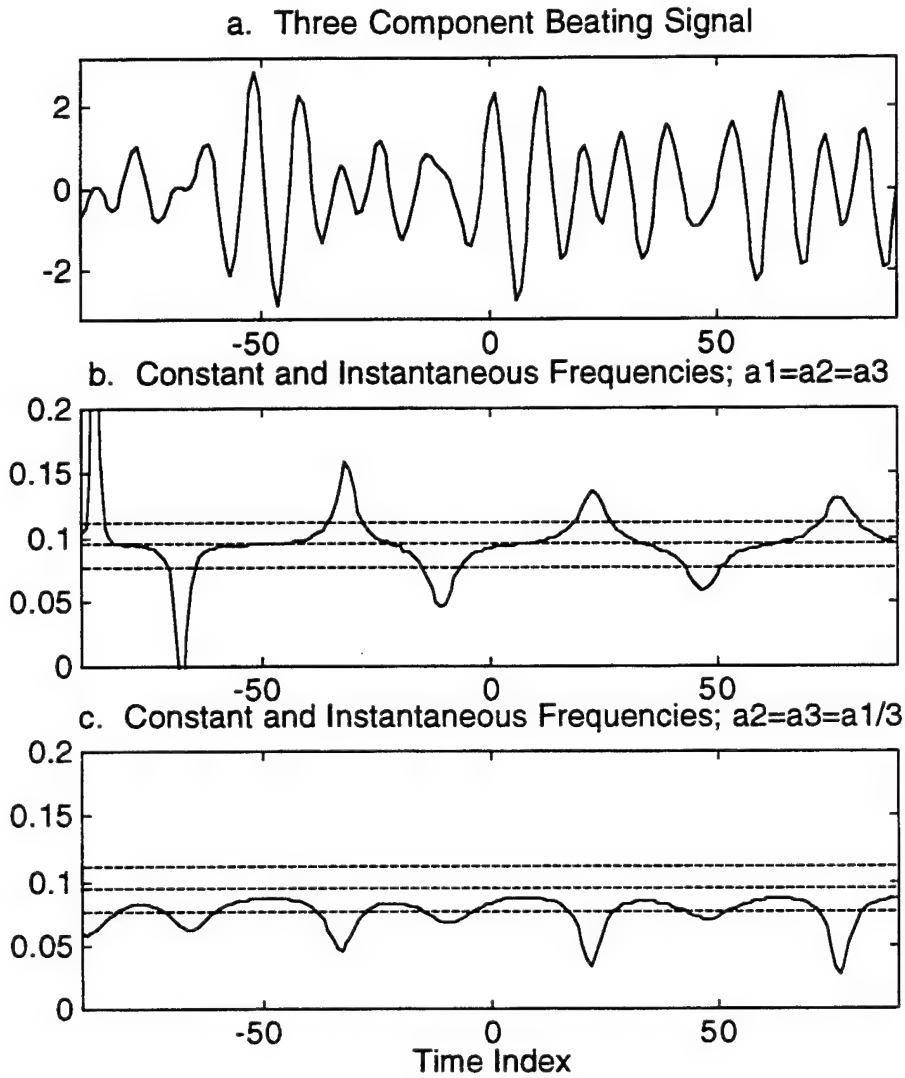

Figure A.4 Illustration of Instantaneous Frequency for Three Sinusoids; a: original signal with equal amplitudes; b: constant ($^{14.314}/_{128}$, $^{12.2}/_{128}$ and $^{9.75}/_{128}$; dashed lines) and instantaneous frequency (solid line) when $a_2=a_3=a_1$; c: constant and instantaneous frequency when $a_2=a_3={}^{a_1}/_3$.

the largest amplitude component (f_1), but that estimate is in fact not symmetrical about that frequency.

An additional useful observation can be made from Figure A.4. By comparing the signal in the upper subfigure to the IF in the middle subfigure, the physical concept of the IF can be clearly seen. Inspection of the signal over time spans of approximately one cycle does reveal variations in the apparent frequency; for example, the period appears to increase for the half cycles just prior to times 0 and 50 and that is reflected in the decrease in the instantaneous frequency over the same time short interval. Further discussion on three-component signals is given in Section A.3.3.

These illustrations (and many additional arguments not presented here) confirm that the concept of "instantaneous frequency", which is popular in the literature in a great many applications, *is not practically useful for interpreting multicomponent signals.* The half cycle (or fraction thereof) phase discontinuity also introduces errors into simple frequency estimation techniques like counting zero crossings, which has been proven to be a sub optimal and biased frequency estimator for signals with a nonstationary IF (Boashash, 1992). However, while direct use of the instantaneous frequency is not recommended, understanding the concept of instantaneous frequency is nonetheless critical because it does greatly improve interpretation of the results from any modeling - especially the new Harmonic Phase Tracking technique.

A.3 Further Algebraic Studies.

Some limited additional topics are listed that were researched but ultimately were not used in the course of this study.

A.3.1. Amplitude Normalization.

Some limited investigations were done in this study to divide the instantaneous signal by the envelope and thereby minimize amplitude variation effects on phase estimation, but they were not completed. For a beating signal this approach does result in an essentially constant amplitude sinusoid. And when the two component amplitudes are similar in magnitude, this resulting "mean frequency" sinusoid does clearly show the 180 degree phase shift at the nodes. However, in more general cases where one amplitude dominates, the 180 degree phase shift in one of the components in Equation A.3 is summed with the complimentary term in that equation that is not experiencing a phase shift, and the "effective" phase shift of the instantaneous signal can be any value between 0 and 180 degrees. This variability in the phase is difficult to confidently detect numerically from the total signal, so the approach was abandoned.

A.3.2. Quadratic Solution.

This was an interesting exercise that originated from Equation A.9, which is repeated here for convenience:

$$x(t) = \left[2\overline{A}\cos(2\pi f^{-}t + \theta^{-})\right]\cos(2\pi f^{+}t + \theta^{+}) + \left[2A_{\Delta}\sin(2\pi f^{-}t + \theta^{-})\right]\sin(2\pi f^{+}t + \theta^{+}) \tag{A.9}$$

The objective of this exercise was to use numerical approximations available from a measured function to find some of the parameters in Equation A.9.

Assume that the interval of the signal under study is long enough to include the maximum and minimum of the envelope (or, multiple values if noise is present so that some averaging can be performed). Then observe that the amplitudes in Equation A.8 can be numerically approximated as:

$$\overline{A} = \frac{\text{mean}(\max(\text{envelope}))}{2} \tag{A.16a}$$

and

$$A_{\Delta} = \frac{\text{mean}(\min(\text{envelope}))}{2} \tag{A.16b}$$

where the mean is intended over multiple cycles of the envelope. Next, use the Hilbert Transform to calculate the analytical signal and then the squared envelope $E^2(t)$ of the signal:

$$E^2(t) = x^2(t) + \mathcal{H}\{x(t)\}^2 \tag{A.17}$$

Calculate the analytical squared envelope defined by the bracketed terms in Equation A.9:

$$E^2(t) = \left(2\overline{A}\right)^2 \left\{ \cos^2\beta(t) + \left({}^{A_\Delta}\!/_{\overline{A}} \right)^2 \sin^2\beta(t) \right\}$$
$$= \left(2\overline{A}\right)^2 \left\{ 1 - \sin^2\beta(t) \left[1 - \left({}^{A_\Delta}\!/_{\overline{A}} \right)^2 \right] \right\} \qquad \text{A.18}$$

where $\beta(t) \equiv 2\pi \bar{f} t + \theta^-$. Equate these numerical and analytical squared envelopes and solve for the $\sin\beta(t)$ term:

$$|\sin\beta(t)| = \sqrt{\frac{1 - \frac{E^2(t)}{(2\overline{A})^2}}{1 - \left(\frac{A_\Delta}{\overline{A}}\right)^2}} \qquad \text{A.19}$$

This result is a [rectified] function of the envelope frequency sinusoid only versus time, so it can be used to find an approximation to the difference frequency (say, using the Instantaneous Frequency approach, and recognize that rectifying doubled the true difference frequency).

Use this estimated difference frequency and the observed envelope phase of θ^- to then estimate the two bracketed envelope components from Equation A.9:

$$E_c(t) = 2\overline{A}\cos\beta(t) \qquad \text{A.20a}$$

and

$$E_s(t) = 2\overline{A}\sin\beta(t) \qquad \text{A.20b}$$

Define the argument proportional to the sum frequency as $\alpha(t) \equiv 2\pi f^{+} t + \bar{\theta}$. Then rearrange Equation A.9 and square:

$$
\begin{aligned}
\left[x(t) - E_c(t)\cos\alpha(t)\right]^2 &= \left[E_s(t)\sin\alpha(t)\right]^2 \\
x^2(t) - 2x(t)E_c(t)\cos\alpha(t) + E_c^2(t)\cos^2\alpha(t) &= E_s^2(t)\sin^2\alpha(t) \\
\left[E_c^2(t) + E_s^2(t)\right]\cos^2\alpha(t) - 2x(t)E_c(t)\cos\alpha(t) + \left[x^2(t) - E_s^2(t)\right] &= 0 \\
E^2(t)\cos^2\alpha(t) - 2x(t)E_c(t)\cos\alpha(t) + \left[x^2(t) - E_s^2(t)\right] &= 0
\end{aligned}
\qquad \text{A.21}
$$

The final line of Equation A.21 is seen to be a quadratic equation in $\cos\alpha(t)$, which means it can be solved for directly (without rectification) at each time step. As previously discussed, a technique such as Instantaneous Frequency would be used to find the mean frequency from this estimated $\cos\alpha(t)$ function. This technique did work well for paired beating signals, although its validity when noise or a third sinusoid was present was not explored. This approach for estimating the difference and mean frequencies (and hence the two component frequencies) was not adopted for the final implementation of the new modeling technique.

A.3.3 More than Two Sinusoids.

What happens if a third (or more) sinusoidal component is present (as in Figure A.4)?

Since the main thrust of this investigation is parameter estimation of real world signals, it must be expected that more than two components will

regularly occur. Chapter 4 explains how the notch filtering properties of the Fourier Series are used to make the initial estimate of the Harmonic Phase Tracking components present in a multiharmonic signal. There is no guarantee that there will be 1 or at most 2 component sinusoids within the bandwidth of all the Fourier Series frequency bins; in addition, there is inevitably a small amount of leakage of energy from other components at neighboring frequencies. Therefore, it is more reasonable to assume that all of the signals will always have more than two components.

Presenting various algebraic expressions for three or more sinusoids is not instructive. If the objective of such algebra is to show a functional form based on a single cosine and sine term with envelopes as in Equation A.3, then each envelope term would be seen to consist of 6 components proportional to three difference frequencies ($|f_2-f_1|$, $|f_3-f_2|$, $|f_3-f_1|$). Practically speaking, the visual effect of this third sinusoid is to superimpose a second slowly-varying oscillation on the original slowly-varying envelope, with an underlying "instantaneous" signal still proportional to some time-dependent sum frequency value.

So generally, if the assumed model consists of two sinusoids, then the best-fit to a 3 sinusoid signal will return biased answers as to those first two frequencies. The assumption made in the development of this technique is that the amplitude of the third sinusoid is small compared to the other two, so that the bias is not large.

[blank]

APPENDIX B

NUMERICAL STUDY OF ESTIMATED PHASE FOR MONOCHROMATIC SIGNAL

This Appendix further explores the claim made in Section 4.2 that the estimated phase used in Harmonic Phase Tracking (HPT) is unbiased when modeling a single sinusoidal signal. This finding was based in part on the proposed relationship presented in Equation 4.16 which is expanded and rearranged slightly here:

$$\frac{\hat{\theta}}{\theta} = \frac{\left(\hat{b}/\hat{a}\right)}{\left(b/a\right)}$$

$$\stackrel{?}{=} \frac{\left(\hat{b}/b\right)}{\left(\hat{a}/a\right)} \qquad \text{B.1a,b,c}$$

$$= \left[\left\{\frac{\left(S_{2\pi f\Delta\xi} + S_{2\pi f\Sigma\xi}\right)}{\left(1 + S_{4\pi\hat{f}\xi}\right)}\right\}\left\{\frac{\left(1 - S_{4\pi\hat{f}\xi}\right)}{\left(S_{2\pi f\Delta\xi} - S_{2\pi f\Sigma\xi}\right)}\right\}\right]$$

A question mark has been added to Equation B.1b because that is the key step which is evaluated in this Appendix.

Note that Equation B.1.c is based on Equation 4.15 for the expected values of the in- and out-of-phase coefficients, repeated here as:

$$\frac{\hat{a}}{a} = \left(\frac{\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi}}{1 + \mathcal{S}_{4\pi \hat{f} \xi}} \right)$$

$$\frac{\hat{b}}{b} = \left(\frac{\mathcal{S}_{2\pi f_\Delta \xi} - \mathcal{S}_{2\pi f_\Sigma \xi}}{1 - \mathcal{S}_{4\pi \hat{f} \xi}} \right) \qquad \text{B.2a,b}$$

A numerical study was done in MATLAB (MATLAB, 1989) to investigate the behavior of these three expressions for various values of the three parameters that control them: the true frequency, the difference between the true and estimated frequencies, and the integration interval. The values selected for the study were arbitrarily chosen simply to span a range of representative values.

Table B.1 summarizes the $\hat{\theta}/\theta$ findings by numerically examining the bracketed term in Equation B.1d. Ideally, all entries would be 1.0 to confirm that the expression was unbiased and could be confidently used in HPT to iteratively adjust the frequencies. The rows in Table B.1 are a sampling of representative periods defined as follows:

- arbitrarily choose bin numbers of 2,10, 18, 26, and 34 relative to a 128-pt FFT (Nyquist Bin 64) to approximate the span of a wideband signal; and
- assume a unit time step.

This bin number set then corresponds to periods between 1.88 to 32 seconds. The columns in Table B.1 define the fractional difference

between the [unknown] true and estimated bin number for the chosen set; assuming that the estimate is reasonably close to the true value resulted in a selected range between -0.5 and 0.5. The integration interval differs for each subtable and ranges from ± 20 to ± 500 points.

As seen from Table B1, this phase ratio is reasonably close to 1.0 for most cases. This generally confirms the use of Equation B.1d for adjusting frequencies. However, it must be recognized that the simple ratio function of Equation B.1.c is not strictly correct because: (1) it "rectifies" the third and fourth quadrants into the first and second, but most importantly, (2) it may be ill-conditioned when either denominator term or conversely when both estimated amplitudes in Equation B.2 approach zero.

So the next step was to numerically evaluate both coefficient expressions in Equation B2. Table B.2 summarizes the findings for $\hat{a}/a$ in Equation B2.a; Table B.3 summarizes the findings for $\hat{b}/b$ in Equation B2.b. Inspection shows that these expressions obviously do approach zero under certain conditions. This prompted a study to identify which variable combinations of frequencies and integration lengths allow the integrals to go to zero. For example, the estimate for $\hat{a}$ in Equation B.2a goes to zero when the numerator goes to zero. It can be shown that this will occur whenever:

$$
\begin{aligned}
\mathcal{S}_{2\pi f_\Delta \xi} + \mathcal{S}_{2\pi f_\Sigma \xi} &= 0 \\
2\pi f_\Sigma \xi \sin(2\pi f_\Delta \xi) + 2\pi f_\Delta \xi \sin(2\pi f_\Sigma \xi) &= 0 \\
(\hat{f} + f)\sin(2\pi(\hat{f} - f)\xi) + (\hat{f} - f)\sin(2\pi(\hat{f} + f)\xi) &= 0 \\
(f)\sin(2\pi f\xi)\cos(2\pi \hat{f}\xi) - (\hat{f})\cos(2\pi f\xi)\sin(2\pi \hat{f}\xi) &= 0 \\
\tan(2\pi f\xi) - \left(\frac{\hat{f}}{f}\right)\tan(2\pi \hat{f}\xi) &= 0
\end{aligned}
\qquad \text{B.3}
$$

If the three variables satisfy this last equation then $\hat{a}$ will be zero and correspondingly the coefficient estimate will be ill-conditioned.

By inspection, the numerator of Equation B.2b goes to zero whenever

$$
\begin{aligned}
\mathcal{S}_{2\pi f_\Delta \xi} - \mathcal{S}_{2\pi f_\Sigma \xi} &= 0 \\
\tan(2\pi f\xi) - \left(\frac{f}{\hat{f}}\right)\tan(2\pi \hat{f}\xi) &= 0
\end{aligned}
\qquad \text{B.4}
$$

These two latter expressions define the parameter sets which lead to ill-conditioned estimates for their respective coefficients. But Equation B.4 shows a second source of ill-conditioning. Note that the denominator in Equation B.2a was not a concern since $\min(\mathcal{S}_{4\pi \hat{f}\xi})>-0.22$ so the denominator is always greater than 0. But there does appear to be a problem with the denominator of Equation B.2b, which goes to zero as the frequency difference approaches zero and the sinc function asymptotically approaches 1.0. This makes no physical sense, however, since the overall ratio in Equation B.2b must asymptotically approach 1.0 when the estimated frequency approaches the true frequency and the amplitudes match. This observation implies that the proper interpretation of the

bracketed term in Equation B.1 is that it is the product of two inner ratios for $\hat{a}$ and $\hat{b}$ as given in Equation B.2, and the behavior of each inner ratio must be independently evaluated as terms approach zero. In other words, Equation B.1c is correct. Thus, take the limit of Equation B2.b as the estimated and true frequencies agree and therefore f_Δ goes to zero and f_Σ goes to 2f:

$$\lim_{f_\Delta \to 0}\left(\frac{\hat{b}}{b}\right) = \lim_{f_\Delta \to 0}\left(\frac{S_{2\pi f_\Delta \xi} - S_{2\pi f_\Sigma \xi}}{1 - S_{4\pi \hat{f} \xi}}\right)$$

$$= \left(\frac{1 - S_{4\pi f \xi}}{1 - S_{4\pi f \xi}}\right) = 1.0 \qquad \text{B.5a}$$

And similarly for Equation B2.a:

$$\lim_{f_\Delta \to 0}\left(\frac{\hat{a}}{a}\right) = \lim_{f_\Delta \to 0}\left(\frac{S_{2\pi f_\Delta \xi} + S_{2\pi f_\Sigma \xi}}{1 + S_{4\pi \hat{f} \xi}}\right)$$

$$= \left(\frac{1 + S_{4\pi f \xi}}{1 + S_{4\pi f \xi}}\right) = 1.0 \qquad \text{B.5b}$$

What can be concluded from all of these discussions? Equations B.5 show that both coefficient ratios will asymptotically be 1.0 when the estimated frequency approaches the true frequency (as expected). Therefore, the only situation where the phase ratio would be expected to diverge appreciably from 1.0 is when the $\hat{a}$ or $\hat{b}$ numerator integral summations approach zero as shown by the parameter relationships in Equations B.3 and B.4.

These analytical arguments can be directly and more simply checked by comparing coefficient ratios from Tables B.2 and B.3 to Table B.1. For example, take the upper left entry in each subtable c for common integration length 160:

- $(\hat{b}/b)$=0.470 from Table B.3.c, ;
- $(\hat{a}/a)$=0.646 from Table B.2.c, ; and
- $\left(\hat{\theta}/\theta\right)$=0.728 from Table B.1.c.

Using these values, a direct check of Equation B.1.c shows that $^{0.470}/_{0.646}$ = 0.728, which is not surprising since the phase expression is the quotient of the two coefficient expressions. But the analytical investigation was necessary, if for no other reason than to identify the "irregular" frequencies where ill-conditioning is possible.

In summary, these evaluations have confirmed the use of Equation B.1.c as the basis within Harmonic Phase Tracking to adjust frequencies, with the knowledge that at some isolated parameter combinations the adjustments may be ill-conditioned. Parameter combinations are not checked in the present numerical implementation of this new technique (which is described in later sections of Chapter 4 and Appendix C) as not worth the computational expense. Experience shows that this ill-conditioning phenomenon does not significantly affect the final results, although it may explain why some component frequencies occasionally oscillate during the iterations instead of asymptotically heading towards the final value.

Table B1. Bias Errors in the Phase Estimates

B.1.a. Results for integration interval=40

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	1.303	1.121	1.036	1.0	0.981	0.981	1.034
6.4	0.954	0.977	0.992	1.0	1.005	1.008	1.000
3.55	1.026	1.013	1.004	1.0	0.997	0.995	0.999
2.46	0.982	0.991	0.997	1.0	1.002	1.003	1.001
1.88	1.014	1.007	1.002	1.0	0.998	0.997	1.000

B.1.b. Results for integration interval=80

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.761	0.915	0.990	1.0	0.979	0.909	0.813
6.4	0.953	0.982	0.997	1.0	0.997	0.982	0.955
3.55	0.974	0.990	0.998	1.0	0.998	0.990	0.975
2.46	0.982	0.993	0.999	1.0	0.999	0.993	0.982
1.88	0.986	0.995	0.999	1.0	0.999	0.995	0.986

B.1.c. Results for integration interval=160

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.728	1.001	1.045	1.0	0.964	1.013	1.275
6.4	0.945	0.999	1.008	1.0	0.992	1.002	1.055
3.55	0.970	0.999	1.004	1.0	0.996	1.001	1.030
2.46	0.979	0.999	1.003	1.0	0.997	1.001	1.021
1.88	0.984	1.000	1.002	1.0	0.998	1.000	1.016

B.1.d. Results for integration interval=300

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	4.080	0.962	1.055	1.0	1.039	1.078	3.529
6.4	0.874	0.969	1.000	1.0	0.993	1.028	0.946
3.55	0.887	1.004	0.994	1.0	0.996	0.991	0.875
2.46	1.052	1.012	1.000	1.0	1.003	0.989	1.022
1.88	1.065	0.998	1.003	1.0	1.002	1.005	1.074

B.1.e. Results for integration interval=500

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.877	0.101	1.058	1.0	1.035	0.132	0.749
6.4	1.072	1.173	1.001	1.0	0.992	1.113	0.970
3.55	1.013	1.190	0.994	1.0	0.996	1.201	1.036
2.46	0.974	0.941	1.000	1.0	1.003	0.959	1.012
1.88	0.993	0.912	1.003	1.0	1.002	0.908	0.981

B.1.f. Results for integration interval=1000

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.842	0.456	0.906	1.0	1.003	0.609	1.161
6.4	1.032	1.153	1.019	1.0	0.999	1.108	0.968
3.55	0.983	0.924	0.989	1.0	1.000	0.945	1.018
2.46	1.012	1.056	1.007	1.0	1.000	1.040	0.987
1.88	0.991	0.959	0.994	1.0	1.000	0.970	1.010

Table B2. Bias Errors in the Estimated In-Phase Coefficients

B.2.a. Results for integration interval=40

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.844	0.929	0.978	1.0	1.010	1.001	0.948
6.4	0.991	1.001	1.002	1.0	0.996	0.986	0.968
3.55	0.955	0.984	0.996	1.0	1.000	0.992	0.968
2.46	0.976	0.994	1.000	1.0	0.997	0.988	0.968
1.88	0.961	0.987	0.997	1.0	0.999	0.991	0.968

B.2.b. Results for integration interval=80

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.998	0.999	0.998	1.0	1.003	1.003	0.967
6.4	0.896	0.969	0.995	1.0	0.995	0.969	0.895
3.55	0.886	0.965	0.994	1.0	0.994	0.965	0.886
2.46	0.883	0.964	0.994	1.0	0.994	0.964	0.883
1.88	0.881	0.963	0.994	1.0	0.994	0.963	0.881

B.2.c. Results for integration interval=160

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.646	0.846	0.954	1.0	0.993	0.841	0.489
6.4	0.571	0.847	0.971	1.0	0.978	0.846	0.540
3.55	0.564	0.847	0.972	1.0	0.977	0.847	0.547
2.46	0.561	0.847	0.973	1.0	0.976	0.847	0.549
1.88	0.560	0.847	0.973	1.0	0.976	0.847	0.551

B.2.d. Results for integration interval=300

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	-0.020	0.534	0.888	1.0	0.895	0.505	-0.022
6.4	-0.055	0.532	0.912	1.0	0.915	0.516	-0.053
3.55	-0.055	0.522	0.915	1.0	0.914	0.526	-0.055
2.46	-0.050	0.520	0.912	1.0	0.911	0.526	-0.051
1.88	-0.050	0.524	0.911	1.0	0.911	0.522	-0.050

B.2.e. Results for integration interval=500

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	-0.133	0.044	0.746	1.0	0.754	0.043	-0.143
6.4	-0.121	0.022	0.767	1.0	0.770	0.023	-0.127
3.55	-0.124	0.022	0.770	1.0	0.769	0.022	-0.123
2.46	-0.126	0.025	0.767	1.0	0.766	0.024	-0.124
1.88	-0.125	0.025	0.766	1.0	0.766	0.025	-0.126

B.2.f. Results for integration interval=1000

Period (sec)	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	-0.098	-0.033	0.271	1.0	0.258	-0.030	-0.084
6.4	-0.089	-0.022	0.256	1.0	0.259	-0.023	-0.092
3.55	-0.091	-0.025	0.260	1.0	0.258	-0.025	-0.090
2.46	-0.090	-0.023	0.258	1.0	0.259	-0.023	-0.091
1.88	-0.091	-0.024	0.259	1.0	0.258	-0.024	-0.090

Table B3. Bias Errors in the Estimated Out-of-Phase Coefficients

B.3.a. Results for integration interval=40

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	1.100	1.042	1.013	1.0	0.991	0.983	0.980
6.4	0.945	0.978	0.994	1.0	1.001	0.994	0.968
3.55	0.980	0.996	1.001	1.0	0.997	0.988	0.968
2.46	0.959	0.986	0.997	1.0	0.999	0.992	0.968
1.88	0.974	0.993	1.000	1.0	0.998	0.989	0.968

B.3.b. Results for integration interval=80

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.760	0.914	0.988	1.0	0.982	0.912	0.786
6.4	0.854	0.951	0.992	1.0	0.992	0.951	0.855
3.55	0.863	0.955	0.993	1.0	0.993	0.955	0.864
2.46	0.867	0.957	0.993	1.0	0.993	0.957	0.867
1.88	0.869	0.958	0.993	1.0	0.993	0.958	0.869

B.3.c. Results for integration interval=160

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.470	0.847	0.997	1.0	0.957	0.852	0.624
6.4	0.539	0.846	0.978	1.0	0.971	0.848	0.570
3.55	0.546	0.847	0.977	1.0	0.972	0.847	0.563
2.46	0.549	0.847	0.976	1.0	0.973	0.847	0.561
1.88	0.551	0.847	0.976	1.0	0.973	0.847	0.559

B.3.d. Results for integration interval=300

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	-0.081	0.513	0.937	1.0	0.929	0.544	-0.079
6.4	-0.048	0.515	0.912	1.0	0.909	0.531	-0.050
3.55	-0.049	0.525	0.909	1.0	0.910	0.521	-0.048
2.46	-0.053	0.527	0.912	1.0	0.913	0.521	-0.052
1.88	-0.053	0.523	0.913	1.0	0.913	0.525	-0.053

B.3.e. Results for integration interval=500

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	0.117	0.004	0.789	1.0	0.781	0.006	-0.107
6.4	0.129	0.026	0.767	1.0	0.764	0.025	-0.123
3.55	0.126	0.026	0.765	1.0	0.766	0.026	-0.127
2.46	0.123	0.023	0.767	1.0	0.768	0.023	-0.126
1.88	0.124	0.023	0.768	1.0	0.768	0.023	-0.124

B.3.f. Results for integration interval=1000

Period	Estimated - Exact Bin Difference (w.r.t 128-pt segment)						
(sec)	-0.45	-0.25	-0.10	0	0.10	0.25	0.45
32	-0.083	-0.015	0.246	1.0	0.259	-0.018	-0.097
6.4	-0.092	-0.026	0.261	1.0	0.258	-0.025	-0.089
3.55	-0.090	-0.023	0.257	1.0	0.259	-0.023	-0.091
2.46	-0.091	-0.025	0.259	1.0	0.258	-0.024	-0.090
1.88	-0.090	-0.023	0.258	1.0	0.259	-0.024	-0.091

[blank]

APPENDIX C

HARMONIC PHASE TRACKING ALGORITHMS

C.1 Overview

HPT is an iterative methodology that asymptotically converges to either the exact answer (for discrete, multiharmonic signals) or a unique answer for stochastic or time-varying signals. Accordingly, the accuracy of that solution is a function of the convergence criteria. As exemplified by the discussions in Chapter 4 and Appendix A, it is necessary to resort to numerical procedures to estimate the optimum values for many parameters. Therefore, it is important to outline and explain all of the steps used in this present implementation of the Harmonic Phase Tracking method. That is the purpose of this Appendix. Chapter 5 presents numerical validation studies using deterministic signals with known parameters, with and without noise; the text in that chapter also compliments the information here.

As shown in Chapter 5, the Harmonic Phase Tracking technique is capable of successfully identifying essentially arbitrary multiharmonic signals with any number of components and constant and/or time-varying parameter values in noise (within the unavoidable constraints presented by: (1) the time-frequency ambiguity, and (2) the degree of nonstationarity). Because of this applicability to such a wide range of signal characteristics, and the fact that engineers and scientists search for such a variety of phenomena from measured signals, it is not possible in this first study to address the strengths and limitations of this technique relative to every conceivable combination of frequency spacings and bandwidths, relative amplitudes, phases, nonstationarities and noise that might be of interest. Regardless, the goal of the implementation as described in this Appendix was to be as universally applicable (robust) as possible, with an accepted consequence that it is not optimal in a numerical sense for most problems. The dilemma for this Appendix is that fully reporting this breadth of use is impractical. Instead, the chosen strategy is to include as much information as possible in a reasonable number of pages by being as concise as possible (such as referencing equations in the main text), with the hope that any incomplete information is at least enough to alert an interested reader to investigate the potential of this technique for their application. Flow charts and coding are not listed.

C.2 Basic Algorithm Assumptions.

1. The coding and analyses were all performed using MATLAB.

2. Unit time steps were assumed. Bin numbers instead of frequency are emphasized since they represent the "non dimensional" number of cycles relative to the FFT reference length. The final frequencies and periods are readily converted using the actual time step. (However, the word "frequency" is used generically throughout this Appendix to minimize confusion (e.g., "difference frequency" rather than "difference bin number"); these text references typically imply bin number in most of the algorithms.)

3. Non-windowed FFTs were used for three reasons: (1) simplicity, (2) the bandwidth for non windowed (or, equivalently, boxcar or rectangular window) signals is narrower in the immediate vicinity of the peak compared to windowed signals, and (3) the physical interpretation of the results is clearer (e.g., the affect of window tapering on signals with linearly nonstationary amplitudes as presented in Chapter 5 was not investigated).

4. It is assumed that the data is discrete and has been properly sampled and prepared (e.g., antialiasing filters).

5. The default data vector for these discussions is assumed to be a summation of sinusoids of unknown rank, frequencies, amplitudes, phases, and

probability distribution, with additive noise with unknown energy and spectral distribution. Zero mean signals are assumed with no loss of generality. Specific categories of signals and/or noise are defined if relevant to particular algorithm performance.

As shown by the multiharmonic example in Chapter 5, the convergence and accuracy of HPT improves as the length of the data vector increases. That is a fact not worth pursuing here. As with many investigations of identification techniques, the thrust of this Appendix and in fact much of the main text is directed towards establishing the performance of HPT using short data segments and non ideal signal types. The reader is reminded of this because most of the text in this Appendix is directed towards maximizing performance for these marginal cases, whereas for moderate data lengths many of these "marginal data" problems and uncertainties are not relevant and the technique converges rapidly and without problems.

The ordering of sections in this Appendix loosely follows the chronological sequence used in the technique as summarized in Table C.1:

Section	Topic
C.3	Construction of the **R** transform matrix
C.4	Estimation of the best two harmonics from a row of **R**
C.5	Estimation of the best single harmonic from a row of **R**
C.6	Adjustment of the **R** matrix to find the initial HPT frequency vector estimate $\hat{\mathbf{f}}^{(0)}$
C.7	Estimation of best frequency vector $\hat{\mathbf{f}}$ via the Harmonic Phase Tracking technique: • criteria for component deletion & addition • frequency correction • convergence measures

Table C.1 Appendix C Outline

C.3 Construction of the R transform matrix

The first step in the HPT technique is to select the number of data points of the segment, denoted here as M. This segment length determines the resolution of the final frequency vector estimate relative to the time-frequency ambiguity. If the segment is too short, then the analysis may

only see a small segment of a component envelope (between any two closely-spaced components) and subsequently will not be able to isolate it. If it is too long, then nonstationarity issues dominate. The validity of any particular choice for M can be argued many ways. Theoretically speaking, ocean waves change constantly and only a nonstationary descriptor is valid to apply; that argument is not constructive and is not accepted. Recall, as stated in Chapter 2, that the ocean can be considered to be "relatively stationary" for shorter periods on the order of hours. But that is too long a definition for one segment from a numerical and engineering standpoint (the resolution would be so good that there would be *too many* components to interpret - even if they had invariant frequencies, amplitudes and phases, which is doubtful; plus, ensemble averaging over independent segments is required because the data is stochastic).

Also, recall the example signal investigated in Chapter 5 defined as the sum of two very closely spaced sinusoids where the particular trial segment of length M starts from an envelope node and spans halfway to the first peak of the envelope. Based on this one segment alone, the signal can be interpreted as either a *two-component stationary* signal or a *one-component nonstationary* signal with a [linearly] time-varying amplitude. There is no way to resolve this ambiguity for such a short signal. And, most importantly, based solely on that limited amount of data *both interpretations are correct.* This also applies to HPT, which may converge to either interpretation depending on the information in the segment.

So is there a defendable strategy for selecting a number of points for a HPT analysis? The suggested answer is a guarded yes, because the behavior of the narrow-band filtered rows in **R** offer the analyst clues regarding the best definition of the segment lengths to be analyzed. At each of these rows the envelope or the transform can be determined and inspected. Note, however, that the first row will always have one cycle regardless of the length of the FFT (N) and will accordingly always be statistically unreliable; conversely, the Nyquist bin row will have N/2 cycles and could be nonstationary over that many cycles. So it must be accepted for geophysical data that for non-narrowband data *every* choice for segment length will not be optimum for all frequencies.

In the HPT iterations for the frequency vector, the maximum forward and backward time shift is presently arbitrarily defined as one quarter of the segment length. Thus, *the total amount of data used to fit the model is equal to the length of the segment of interest plus one quarter length before and after = 1.5*M.* All of this data is used to construct **R**, so rows have a length of 1.5*M-N.

The chosen strategy for selecting the "optimum" segment length (M) is to examine the behavior of the row of **R** which has the most energy (peak bin number of the spectrum). Assume that multiple components are present due to the finite resolution (proportional to the length of the FFT

used to construct **R**). *The suggested approach is to select the segment length such that the row length of **R** is long enough to identify at least one, or better two cycles of the "dominant" component of the envelope.* This is enough to allow for reliable component resolution, so in the general case there is no need to make it longer and potentially encounter nonstationary effects. This is also consistent with the modeling assumption of no more than two sinusoids to fit the [arbitrary] envelope at each row of **R** as discussed in Chapter 4.

So, the choices for segment length and FFT length are *necessarily* subjective, even for a given data vector. The strategy used for these present studies is summarized as:

1. Choose a FFT length N such that the bandwidth of the spectrum includes at least a dozen or so bins.

2. Choose a trial length of test segment M and construct **R** by appending the real column transforms in chronological order. Each segment is defined by shifting the data vector forward by one step.

3. Identify the row/bin number with the most energy, denoted k_{peak}.

4. Adjust M so that the row k_{peak} of **R** (in MATLAB notation $R(k_{peak},:)$) has nominally two crests, or more if the crests appear consistent (implying a low number of stationary sinusoids in that bin).

C.4 Estimation of the best two components from a row of R

This set of routines is also used to insert new frequencies into the frequency vector during the iterations, so it is an important workhorse routine of this new technique. The time series can either be a row of **R** or a residual error vector defined as the difference between the signal and an intermediate best-HPT-fit during the iterations. It also has to have the most robust algorithms, since the vectors it sees can and generally do have characteristics ranging from deterministic to practically random. To maximize their applicability to all types of data vectors, most of the algorithms are numerical (the problems associated with the [improper but popular] use of instantaneous frequency and/or counting zero crossings as described in Appendix A are two good examples of why analytical or simple routines are not robust).

This step of identifying two components is done in three stages. The general approach is to: (1) numerically estimate the difference and average frequencies; (2) algebraically convert those two frequencies to the two estimated component frequencies via Equation A.2; and (3) numerically iterate to improve those estimates based on least squares fits to the data vector.

The first stage of estimating the difference (envelope) frequency is difficult for several reasons: (1) the numerical envelope is a positive/rectified function which induces biases; (2) it must be smoothed to partially remove dynamic components from higher order components, further adding to the bias; and (3) there may be only one or even less cycles to fit when the two sinusoids are near the minimum frequency spacing allowed by the method.

First, the envelope is calculated using the Hilbert Transform. This estimate is low pass filtered for three reasons: (1) to minimize possible severe numerical end effects; (2) minimize noise and higher frequency components due to leakage interactions with sinusoids outside the subject bin (these other sinusoids will be identified later when their particular rows are examined); and (3) emphasize closely-spaced components located in the subject bin. The resulting signal is the best, smoothed and rectified representation of the original dynamic envelope from **R**.

HPT assumes that this envelope is due to at most *two* underlying harmonics such that the zero-mean (not rectified) envelope function is itself a constant parameter single harmonic. That makes the objective of the next step to fit a sinusoid to the smoothed, rectified envelope signal (which appears as a sinusoid with twice the frequency of the zero-mean envelope). There can be two complications: (1) in many cases, particularly for low bin numbers, there is only between a half and one cycle available; and (2) a beating envelope with zero-valued nodes has a

discontinuous derivative if the node is zero valued (because of rectifying), so fitting a sinusoid to the envelope such that the sinusoid trough equivalently fits this node introduces a bias in the fit (which is not considered significant when the objective is only to fit the frequency).

The following algorithm evolved to recover the rectified frequency of this smoothed envelope signal:

1. define a unit length, zero-mean, normalized time vector, equal in length to the length of the envelope, and with maximum values of $\pm 1/2$

2. algebraically decompose the signal into (even) in- and (odd) out-of-phase components (e.g., even component is only cosine and odd is only sine relative to the center time)

3. estimate the two best fit in- and out-of-phase amplitudes (e.g., cosine crest versus cosine trough) [this can be very ambiguous and difficult]

4. fit polynomials in powers of time to the even and odd functions, and select the order of fit (say, 10) that provides a good match without exhibiting diverging behavior for the large time values (this explains the need for normalizing the time vector in the first step); then, fit a polynomial to the [rectified] total envelope e(t).

5. Recall the series expansions for sine and cosine at frequency ω:

$$\cos(\omega t) = 1 - \frac{(\omega t)^2}{2!} + \frac{(\omega t)^4}{4!} - \cdots$$

$$\sin(\omega t) = (\omega t) - \frac{(\omega t)^3}{3!} + \frac{(\omega t)^5}{5!} - \cdots \qquad \text{C.1a,b}$$

6. Group the polynomial terms for the total envelope e(t) from step 4:

$$\begin{aligned}\hat{e}(t) &= a_0 + a_1 t + a_2 t^2 + a_3 t^3 + a_4 t^4 + a_5 t^5 \cdots \\ &= \left[a_0 + a_2 t^2 + a_4 t^4 + \cdots\right] + \left[a_1 t + a_3 t^3 + a_5 t^5 \cdots\right] \\ &= a_{even}\left[1 + \left(a_2 / a_{even}\right) t^2 + \left(a_4 / a_{even}\right) t^4 + \cdots\right] + \\ &\quad a_{odd}\left[\left(a_1 / a_{odd}\right) t + \left(a_3 / a_{odd}\right) t^3 + \left(a_5 / a_{odd}\right) t^5 \cdots\right]\end{aligned} \qquad \text{C.2}$$

where $a_{even} \equiv a_0$, but $a_{odd} \equiv a_1$ is more difficult to estimate.

7. Equate Equation C.1a with the first bracketed term in Equation C.2, and equate Equation C.1b with the second bracketed term in Equation C.2, and note that for each term:

$$\left|f^{(j)}\right| = \left|\frac{\left(j!\, a_j / a_{max}\right)^{1/j}}{2\omega}\right| \qquad \text{C.3}$$

where a_{max}=appropriate cosine or sine amplitude (a_{even} or a_{odd})from step 3, and $f^{(j)}$is an estimate of the frequency of the rectified envelope from the t^j-power coefficient. Since there are P-1 frequency estimates from the Pth-order polynomial, the best frequency $\bar{f}_{env}$ is taken as an amplitude weighted average of all of them (e.g., if signal is primarily a cosine, then the sine component will be negligible and the associated frequencies will be unreliable). Note that $\bar{f}_{env}$ is the mean, rectified envelope frequency; it is not yet the zero-mean envelope frequency.

8. Find least squares estimates of a_{even} or a_{odd} using this weighted $\bar{f}_{env}$; re-estimate $\bar{f}_{env}$ using Equation C.3, and iterate until the coefficients and $\bar{f}_{env}$ converge (that is, the amplitudes and frequency in Equation C.3 are consistent).

9. Perform a subsequent numerical trial search to refine this frequency estimate by using a least squares fit directly between a modeled sine and cosine and the rectified envelope.

10. The estimated frequency of the zero mean envelope (difference frequency) f_Δ is then half of this refined frequency.

This seems like a very awkward technique, but experience showed that simpler schemes could be easily fooled when the data consisted of only part of an envelope cycle with large additive noise, or had an envelope with more than two components evident.

Stage 2 of this first task is to estimate the average frequency of the instantaneous signal defined as the kth row of **R**. Since analytical techniques like instantaneous frequency and counting zero crossings are unreliable (see Appendix A), the chosen strategy was to simply define the first estimate of this average frequency as the mean integer value for the subject bin. Trial component estimates of $\hat{f}_1^{(0)}$ and $\hat{f}_2^{(0)}$ are then found using Equation A.2.

The third and final stage is to iterate these initial component estimates to find the best two harmonics that fit the subject row of **R** (or residual data vector in the component insertion step of the technique). This is essentially a numerical optimization process that defines the best frequency pair as the one that yields the smallest least squares error $Q\left(f_1, f_2\right)$, or equivalently, $Q\left(\bar{f}, f_\Delta\right)$, where the frequencies are defined in Equation A.2.

The process proceeds as follows:

1. Observe that $\left.\frac{\partial Q\left(\bar{f}, f_\Delta\right)}{\partial \bar{f}}\right|_{\min\left(Q\left(\bar{f}, f_\Delta\right)\right)} >> \left.\frac{\partial Q\left(\bar{f}, f_\Delta\right)}{\partial f_\Delta}\right|_{\min\left(Q\left(\bar{f}, f_\Delta\right)\right)}$; that is, the least squares error is much more sensitive to errors in the average frequency than it is to comparable errors in the difference frequency (which affects the modulating envelope only). Since the derivative with respect to f_Δ is relatively small, it is reasonable to expect that the frequency pair found by holding f_Δ constant and estimating $\bar{f}$ will be very close to the optimum. So, calculate $Q\left(\bar{f}, f_\Delta\right)$ for a finite set of f_1 and f_2 for all values of $\bar{f}$ between $\left(k-1/2\right)/N < \bar{f} < \left(k+1/2\right)/N$ (recall N is the FFT length) such that f_Δ from the first stage is constant. [This searches one row of $Q\left(\bar{f}, f_\Delta\right)$.]

2. Next, calculate $Q\left(\bar{f}, f_\Delta\right)$ for a finite set of f_1 and f_2 for a range of f_Δ such that $\bar{f}$ is kept constant. [This searches one column of $Q\left(\bar{f}, f_\Delta\right)$.]

3. Finish with two final localized searches by varying f_1 with f_2 constant then f_2 with f_1 constant. These last searches are only marginally effective but are done more for completeness here (since $\bar{f} = \left(f_1 + f_2\right)/2$ only when the two amplitudes are equal, so for most applications this last correction is justified).

The best values for the two amplitudes and two phases are taken from the least squares fit using the two frequencies where $Q\left(\bar{f}, f_\Delta\right)$ is a minimum.

In the case of fitting two sinusoids to rows of **R**, the rows are chosen according to those that have the maximum energy. This is an important point because it minimizes leakage from adjacent rows. Therefore, if the process above finds: (1) either component frequency well outside of the bin centered at the integer bin number of interest, or (2) either component very close to a neighboring integer value, then that frequency set is discarded as unrealizable (or more accurately, better estimated when another row of **R** is fitted), and a one sinusoid best fitting is performed. This is described next.

C.5 Estimation of the best one component from a row of R

The last paragraph gave a condition where a one sinusoid fit is used in place of the two sinusoid fit. But the single sinusoid fit is also obviously appropriate when there is only one true sinusoid in any particular bin.

In those cases, if the normalized difference between the maximum and minimum of the [rectified] envelope for the subject row are is less than some threshold (say, 10 percent of the mean envelope value), then that variation is assumed to be due to noise or numerical end effects from the Hilbert Transform. The conclusion is made that there is no "component" beating and that therefore only one significant true sinusoid is present in that bin. In these cases a one-dimensional search is conducted to find the optimum frequency, and with it the optimum amplitude and phase. Care must be taken to avoid a small set of points at each end of the envelope where there may be spurious numerical end effects.

C.6 Adjustment of the **R** matrix to find the initial frequency vector estimate $\hat{f}^{(0)}$

The logic for this process was described in Section 4.7, along with some of the numerical complications in converting the frequency domain amplitude and phase into the time domain. The focus of this section is on the calculation of the component $\mathbf{R}_1$ and $\mathbf{R}_2$ matrices corresponding to the two estimated harmonics.

The easiest way to calculate these matrices would be to define the time domain signal using the time domain amplitude, frequency and phase, and a time vector defined as t'=[0,1 ... N-1], then use an N-point FFT to calculate one column of $\mathbf{R}_j$. Then, shift the time vector by one point and find the

transform (1.5*M-N-1) more times and calculate the remaining columns. This direct method would be very CPU intensive and is not the optimum choice.

The first realization that construction of these matrices could be simplified came from recognizing that since each matrix represents only one sinusoid shifted in time, then the pattern for each row (i.e., in the column space) will be identically sinusoidal (as described text in Section 4.7 and exemplified by Equation 4.46), and that both matrices would be rank 1. However, the series nature of the FFT creates end discontinuities for non-integer-period (i.e., fractional) sinusoids, and the resulting transforms have an unknown bias due to aliasing. This prompted application of singular value decomposition to a variety of representative matrices to determine typical ranks and singular values for these biased matrices. If there were only a small number of singular values, then it would eliminate many of the calculations for the sinusoids and the FFTs. Indeed, decomposition showed that for most values of fractional frequency, phase, and FFT length, the matrices were essentially rank one, meaning that one FFT and one sinusoid were sufficient to construct the matrices (i.e, the shape of the transform was invariant as it was scaled up and down by the time-induced phase variation).

The next step is calculation of each $\mathbf{R}_j$ component matrix corresponding to the two identified harmonics. Recall that the two harmonics were fitted to

the [frequency domain] rows of **R**. The present algorithm calculates the transform of the jth harmonic to define how that fractional frequency would have been distributed among the FFT harmonics used to construct the original **R** matrix; i.e., the real part of this transform is the first column of $\mathbf{R}_j$. To expand this first column into the rank-1 $\mathbf{R}_j$ matrix, this column is simply scaled by a properly phased sinusoid to make each subsequent column. Thus, $\mathbf{R}_j$ is the best-estimate of what the **R** matrix would be if the time domain signal was defined by only that harmonic. The final step is to calculate a "corrected" matrix $\mathbf{R}-\mathbf{R}_1-\mathbf{R}_2$. The process repeats by finding the new row number with the largest variance and recorrecting the **R** matrix until the remaining largest variance reaches a convergence threshold.

This simplicity of generating the rank-1 $\mathbf{R}_j$ component matrices is only violated in two cases. The first is when the sinusoid is in one of the first few bins; these matrices require more than a rank one fit, so for those cases the direct calculation method must be used. The second case applies to those sinusoids where the combination of frequency and the initial phase results in a negligible real transform at all frequencies over the first segment. The modified solution adds an artificial shift of approximately one quarter period to the time vector, calculates the transform, then uses that as the characteristic column vector.

C.7 Estimation of best frequency vector $\hat{\mathbf{f}}$ via the Harmonic Phase Tracking technique.

This section of the MATLAB coding performs the iterative frequency updating that is the heart of this new technique. As outlined in Table C.1, the coding can be loosely organized into 3 categories: (1) criteria for component deletion & addition; (2) frequency adjustment via phase tracking; and (3) convergence measures.

Criteria for component deletion & addition are listed first because they are used to condition the initial estimated frequency vector $\hat{\mathbf{f}}^{(0)}$ that is assembled during the **R** adjustment process. These criteria are checked at every iteration of the frequency updating. There are four criteria for deleting frequencies and three criteria for adding frequencies.

With respect to deleting frequencies (or equivalently bins, which are used in the coding), the first obvious criterion is the difference between adjacent bin values. If this is too small, then mathematically the least squares basis matrix condition number increases too large, and physically the envelope becomes so long that it appears as approximately constant over the finite data length. This provides a physical justification for the chosen remedy of replacing two sinusoids that converge closer than the minimum allowable bin difference threshold with integer sinusoids. The value chosen for this threshold was based on numerical studies and the following relationships: define a scalar κ as

$$\kappa \bullet \text{Length of data segment} \equiv \text{Period}\Big|_{\text{longest resolvable envelope}}$$

$$= {}^{1}\!/_{\min\,(\text{frequency difference})} \qquad \text{C.4}$$

$$= {}^{N_{FFT}}\!/_{\min(\text{bin difference})}$$

Equate the first and last lines and solve for the minimum bin difference to use as the resolution limit for bin deletion:

$$\text{minimum bin difference} = {}^{N_{FFT}}\!/_{(\kappa \bullet \text{Length of data segment})} \qquad \text{C.5}$$

From overall numerical results, a conservative value for κ was found to be approximately 1.6, meaning that HPT could reliably model two beating sinusoids spaced only 0.63 relative bin numbers apart. Note that $\kappa \equiv 1$ for traditional Fourier analyses, so by this measure HPT has 37 percent better resolution than traditional spectral analysis. This is not a surprise since the Fourier frequencies are defined solely on the basis of orthogonality, with no explicit claims that they represent the maximum possible resolution. Also, note that 0.63 cycles corresponds to 226 degrees of a sinusoid. Observations from the many numerical studies show that this choice of κ limits condition numbers of the total least squares basis matrices to less than 10 (for up to rank 70 matrices).

The second deletion criterion is for bin numbers that adjusted too low; an absolute minimum bin threshold of 0.67 was defined. This was considered safely consistent with the 0.6 cycle bin resolution threshold, and no further investigations were done to relate them.

The third and fourth criteria for component deletion involve the component amplitudes. The third criterion detects when two bin numbers are approaching the minimum relative bin spacing, but their respective amplitudes are unreasonably large compared to the local spectral amplitude (using the standard definition of the square root of [two times the FFT spectral ordinate times the frequency bin width]). This criterion can be triggered when two beating sinusoids are phased such that there is a node of the envelope near the center of the test segment; when the envelope period is very long the small segment of envelope appears to grow essentially linearly. In these cases there is an infinite number of combinations of long periods (corresponding to very large κ values) and associated (large) amplitudes that will reasonably fit the small segment, and the total least squares algorithms sometimes estimate physically unrealizable amplitudes. In fact, this is a situation where the total least squares solution is vastly superior to the conventional least squares solution. In these situations, if the bin numbers are approaching the resolution threshold and if the two trial amplitudes exceed twice the maximum adjacent FFT-derived amplitude, the two sinusoids are combined.

The fourth set of criteria for deletions is done simply for computational efficiency. When a component amplitude becomes negligible (because it corresponds to an incorrect initial frequency bin while the remaining frequency estimates iteratively approach their true values), it ceases to contribute and is removed from the frequency vector. Two criteria

defining "negligibility" are used: (1) component amplitudes less than $^1/_{25}$ of the maximum FFT-spectral amplitude (absolute criterion), or (2) less than $^1/_3$ of the smallest of the three most adjacent FFT-spectral amplitudes (relative criterion). This latter check is important because it is often of engineering interest to detect superharmonics located at integer multiples of some fundamental (fractional) frequency, but their amplitudes are often small and the use of one absolute criterion could eliminate them as insignificant.

The first two criteria for adding components are based on absolute and relative error checks found from FFT-based spectra of the original data vector S_{xx} and the residual error S_{ee} (the residual error is defined as the difference between the original data and the fitted data) found at the end of each iteration. Identifying logic to detect when to insert components was one of the most difficult steps in implementing this technique. There are so many combinations of signal and noise vectors that logic to insert in one case would either not insert or even worse insert ad infinitum in other cases. The final implementation as described below was found to be fairly dependable but it is by no means considered rigorous nor optimum.

Secondly, components are always added in pairs. The same algorithms used to fit the rows of the **R** matrix are used, except that here the time domain residual error vector is fitted rather than the row vectors of **R** used before.

The absolute threshold is the first criterion for defining bin insertions and it is adjustable during the iterations. It was found necessary to restrict the definition of the error to a [generous but] finite bandwidth loosely centered around the peak of the FFT-spectrum of the original data. [This avoids adding components for cases such as one sinusoid in low white noise; even an accurate rank one fit would show a significant residual error with an rms value proportional to the sum of the white noise spectral ordinates from 0 to Nyquist bins. But inserting many small components over all bins to reduce this wideband noise would not be an efficient way to model the data. Use of the finite bandwidth minimizes detection of this type of rank inefficient solution.] The threshold ratio of error to original spectral areas (i.e., variances) that triggers component insertions is initialized to 0.5 percent. However, this value might be unattainable for data with a low signal to noise ratio (SNR); accordingly, if early insertions are triggered too often (say every other iteration), this absolute error threshold is dynamically relaxed.

The relative error threshold is the second criterion and is triggered whenever a spectral ordinate from the residual signal at a particular bin k, $S_{ee}(f_k)$, is significant compared to the corresponding ordinate of the data vector $S_{xx}(f_k)$. But insertion based on this relative error is again restricted to a finite bandwidth to avoid unreliable noise-dominated regions of the spectrum. Within the main spectral bandwidth this relative threshold weighs not only the local spectral ordinate ratio $S_{ee}(f_k)/S_{xx}(f_k)$

but also whether $S_{xx}(f_k)$ is significant compared to mean($S_{xx}(f)$, f within bandwidth).

Two scenarios are given to illustrate the need for both of these criterion. The need for the absolute error threshold is exemplified by data with a unimodal spectrum where the error is moderately large but reasonably uniformly distributed in frequency across bandwidth such that none of the individual error spectral ordinates are large enough to trigger the relative threshold. Conversely, the intent of the relative error threshold is to insert components in cases where the absolute (rms) residual error is relatively small but the local residual error is large at a bin of interest where the local spectral energy is small compared to the peak spectral energy, as in the case of modeling smaller amplitude superharmonics.

The third criterion for adding components applies to some cases after two bins are merged based on the minimum allowable bin spacing. Occasionally, the residual error shows large increases afterwards, indicating that combining operation removed components that were actually major contributors to the signal but were skewed due to incorrect off-diagonal terms in the basis matrices. This situation can occur whenever two true components are spaced approximately at the minimum bin threshold as defined by the [trial] choice of the length of time series being analyzed. Thus, whenever the absolute error shows a large increase between iterations (say, 3 percent increase), a new pair of components is added.

In all of the insertion situations, the bin number corresponding to either the peak of the residual error spectrum (for the absolute threshold and residual increase criteria) or the local bin number (for relative criterion) is identified. This defines the center band for a notch filter used to localize the residual data vector before being passed to the addition algorithms.

There is one last set of thresholds to discuss that relate to these bin insertion criteria, and that is local bin difference thresholds. The third situation for component insertions just described (triggered by combining two closely-spaced components which results in a large increase in the residual error such that two new components are inserted) can occur ad infinitum when two true signal components are present just below the minimum detection bin spacing of the technique; the two new components are added, they quickly converge towards the true component frequencies, then are removed when they come too close together, so the cycle repeats. The pattern of bin insertions is checked at each iteration, and if any bin number appears too often, then the minimum bin spacing used as the threshold to combine components at that bin is reduced (to, say, 90 percent of the default value), and the iterations are continued. This reduction is done up to three times at any bin, and is terminated if the threshold is reduced to 70 percent of the original bin threshold (defined in Equation C.5. (This reduction process is seen to adjust the κ value in Equation C.5 even smaller than the default bin resolution value.)

This reduction is successful in most cases. But if the true signals are spaced just below this 70 percent reduced bin value, this combining-inserting pattern will still repeat ad infinitum. In these situations another strategy is required. Since this criterion reduction process cannot be continued beyond physical realizability to where the bin separation is very small (and subsequent problems due to an increased condition number of the total least squares basis matrix), the only alternative is to relax the local spectral relative criterion used to trigger the bin insertions. This relaxation can be performed as often as necessary (with each relaxation defined as a 10 percent reduction for example).

This final step of relaxing the local criterion at that bin is always capable of stabilizing the iterative process, but both of these bin spacing and error criterion adjustments come at the obvious costs of: (1) increasing the number of iterations, and/or (2) increasing the error in the final solution. The first one cannot be avoided with this technique (but it may be possible to minimize it for particular data characteristics). The increased error is not as large a problem, since inspection of the final results will clearly show that the residual error is relatively large at some bins, which alerts the analyst to increase the length of the test segment (M) to increase frequency resolution and repeat the analysis.

As presented in Chapter 4, the key to HPT is shifting the time series forwards and backwards and estimating component phases for each

segment. Some details of the numerical implementation of the second of the three 3 categories of algorithms listed in this section for accomplishing that, namely, frequency adjustment via phase tracking, are presented next.

1. The segment length M is held constant.

2. Since the frequency vector is constant for all shifted segments, one basis matrix is calculated similar to Equation 4.30.

3. Seventeen time shifts (eight forward and eight backward from the center segment) are used. The maximum shift is arbitrarily set equal to one quarter the length of the center segment, so max(shift) = M/4. Non uniform shifts are used to accomodate short and long period components (see step 5).

4. The nominal (and maximum) number of cycles used for estimating the frequency at each bin was set at eight. The reason is that all of the estimated phases (radians) are modulo(2π), and they must be unwrapped to provide the linear slope needed to fit the frequency. This unwrapping is done numerically rather than analytically for the following reason. If the true frequency was known, then the estimated (wrapped) phases versus time shift could be analytically unwrapped; since the true frequency is unknown, this is not possible. Instead, the phase differences are checked starting from the center time and working outward; if a difference between successive phase estimates is greater than π then 2π radians are

added to that phase *and* all of the larger time shift phases; the center time is indexed outward and the checking continues. But it is possible that any numerical scheme like this may be fooled if significant noise is present or if it is carried too far and errors accumulate; so for that reason, this unwrapping process is conservatively applied to no more than four cycles forward and four backward. Also, note that there are less than eight cycles for low bin number components; for example, the plus and minus one quarter segment maximum shift only provides one half cycle of phase information for a component at bin number 1. This defines the total number of phases used for the slope estimation in step 7.

5. Note that both small and large time shifts are required. The periods corresponding to the highest frequency components are on the order of only a few time steps, so small time shifts are necessary to provide a reasonable number of phase estimates for their slope estimation. Conversely, large time shifts are necessary to span the periods corresponding to the lowest frequency components. Therefore, an unequal time shift vector is used to provide both closely-spaced and widely-spaced phase estimates.

6. For each segment, the right hand side vectors such as shown in Equation 4.30 are numerically evaluated following Equation 4.31. An estimated phase vector for that segment is then computed using total least squares and Equation 4.38.

7. After all segments have been analyzed, the phase function at each component frequency is analyzed. First, the function is truncated (if necessary) to limit the span to eight total cycles. Then the best linear slope is fitted. [If the variance of the error defined as the phase vector minus the linear fit is large (say, more than $\pi/2$), then a second check is made to see if one or more points need a quadrant correction and a new slope may be fitted.]

8. The apparent updated frequency is then simply the previous frequency multiplied by the ratio of this new slope to the slope from the previous iteration.

9. This apparent change based on the slope ratio is next modified by several functions per component to improve the stability of the iterative process. For example, a scalar coefficient of determination R^2 (Montgomery and Peck, 1992) is calculated using:

$$R^2 \equiv \frac{SS_R}{SS_R + SS_E} \qquad \text{C.6}$$

where SS_R is the regressor sum of squares and SS_E is the error sum of squares, both defined by:

$$SS_R = \sum_{i=1}^{n} \left(\hat{\Omega}_i - \overline{\Omega}\right)^2$$

$$SS_E = \sum_{i=1}^{n} \left(\Omega_i - \hat{\Omega}_i\right)^2 \qquad \text{C.7}$$

Ω_i are the data-derived phase estimates, $\hat{\Omega}_i$ are the best linear fitted phases, and $\overline{\Omega}$ is the estimated mean (all unwrapped). In words, SS_E measures the variance error between the estimated and fitted phases while SS_R measures the variance in the model due solely to the linear variation. Note that $0 \leq R^2 \leq 1$, with 1 corresponding to deterministic phases that fall exactly on a linear slope (i.e., $SS_E=0$), and 0 corresponding to a zero slope with any error in the data (i.e., $SS_R=0$). As applied here, a large R^2 value implies that the frequency needs consistent updating for all (shifted) segments; for these cases the apparent updating is increased to accelerate the iterative process. Alternatively, a very small R^2 value denotes that there is no linear trend, indicating that the previous and updated slopes and hence the frequencies are converged (at least relative to the basis matrix used during that particular iteration); for these cases the slope ratio is decreased to minimize possible oscillations. As a second example of how the apparent change is modified, recall how the minimum bin spacing can be reduced when the technique detects repeated insertions in one bin region, implying trouble with the iterative controls in that bin region. So, all apparent frequency updates are reduced further in these regions to minimize potential over-corrections that could falsely trigger further bin combinations and associated insertions. A third check reduces this change even further if the

subject bin has a neighboring bin close to the minimum allowable bin difference.

10. All of the modified, apparent updated frequencies are compared to an absolute maximum and minimum allowable relative change; if the apparent change is too large then it is redefined as the maximum allowable value. This check was added to insure stability for problems with large rank. [Of course, the value for this maximum change can be set high or the check can even be eliminated if desired.]

Note that these modifications and limitations only act to accelerate, slow down, or generally stabilize the convergence process. They do not bias the direction of the frequency updates or the final solution in any way.

The final set of algorithms to discuss in this subsection are the convergence measures. Ideally, the measure of convergence of the solution is simple: (1) unchanging rank, with (2) all parameters (frequencies, amplitudes, and phases) converged to any specified number of significant digits. This is sometimes unattainable when all of the various numerical effects are considered.

So instead the questions become: (1) is it possible to construct a "cost" function that trades off rank with residual error, (2) what thresholds for

convergence of frequency, amplitude and phase are attainable, and (3) which thresholds can be invariant and which must be adaptive during the iteration process? These questions are addressed next.

The first issue to address is rank. This, in turn, requires identification of the analyst's objective(s) for analyzing a time series. These objectives can be quite different. For example, an applied mathematician or strict signal analyst is typically interested in either: (1) fitting the time domain signal as exactly as possible, or (2) interpreting the time series as information, and then identifying the minimal rank over all possible fits that conveys the maximum amount of information. Both of these objectives can be and are conveniently quantified and used as an absolute measure of the "efficiency" of a model. For engineers the objective is often quite different. Their main objective is typically to understand the phenomenon by interpreting a time series both qualitatively and quantitatively, and it is the qualitative aspects that introduce the difficulties. For example, an engineer studying the response of a nonlinear dynamic system (such as a building or ocean waves traversing into shallow water) may be interested in identifying whether a particular super harmonic is present or not. Since superharmonics usually have a much smaller amplitude than the fundamental, they can be easily discarded by a signal analyst as not making a significant contribution to the overall fit. The implementation chosen here is that no effort was made to apply any measures to the rank of the fit in this technique (other than

to eliminate components with very small amplitudes as discussed previously in this section), and to instead present all remaining components for consideration. Convergence requires constant rank for at least 10 iterations.

The second question of setting convergence thresholds must be addressed. This is an important issue because the technique is iterative, and even in well-behaved cases the iterations must be truncated as "acceptable" at some point. It is even more important for signals with large noise where the parameter errors may not only be relatively large but also where the lack of any physical (true) sinusoids to track, and/or coupling between components, may make the iterative process prone to wandering instead of monotonically converging. This section only presents an overview of the convergence checking process.

In overview, convergence checks are initiated only after the rank has been constant for at least six iterations. Second, both scalar and vector convergence checks are used. Third, both relative and absolute convergence thresholds are used.

The primary convergence checks involve the frequency, amplitude and phase vectors versus iteration. The variability in the amplitudes and phases are measured by the absolute value of the maximum minus the minimum parameter values at each frequency over the last six iterations.

The variability in the frequency vector estimate is measured by the coefficient of determination for only the subject iteration as defined in Equation C.6. All of these three parameter vectors are next weighted by a normalizing vector based on the component amplitudes. This normalizing is important because a small amplitude component can exhibit relatively large amplitude and phase variations due simply to small adjustments in the amplitudes of large, neighboring components through coupling in the basis matrix; normalizing the convergence measures de-emphasizes these relatively large but in reality small absolute variabilities in the iterative process down to an acceptable "weighted" measure of variation.

Identifying quantitative values used to define convergence of these vectors was difficult, particularly when the goal was to be as generally applicable as possible. Besides the previously-described problem of anticipating what types of measures are important, the fact that the convergence can behave asymptotically means that small reductions in these quantitative thresholds can greatly increase the number of iterations. The chosen values represent reasonable compromises between accuracy and computational resources.